Hindawi
Mathematical Problems in Engineering
Volume 2023, Article ID 9794835, 1 page
https://doi.org/10.1155/2023/9794835

Retraction

Retracted: Visual Object Tracking Based on Deep Neural Network

Mathematical Problems in Engineering

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Z. Diao and F. Sun, "Visual Object Tracking Based on Deep Neural Network," *Mathematical Problems in Engineering*, vol. 2022, Article ID 2154463, 9 pages, 2022.

Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2154463, 9 pages
https://doi.org/10.1155/2022/2154463

Research Article

Visual Object Tracking Based on Deep Neural Network

Zhifeng Diao[1] and Fanglei Sun[2]

[1]*College of Design and Innovation, Tongji University, Shanghai, China*
[2]*School of Creativity and Art, ShanghaiTech University, Shanghai, China*

Correspondence should be addressed to Fanglei Sun; sunfl@shanghaitech.edu.cn

Received 17 May 2022; Revised 12 June 2022; Accepted 16 June 2022; Published 12 July 2022

Academic Editor: Naeem Jan

Computer vision systems cannot function without visual target tracking. Intelligent video monitoring, medical treatment, human-computer interaction, and traffic management all stand to benefit greatly from this technology. Although many new algorithms and methods emerge every year, the reality is complex. Targets are often disturbed by factors such as occlusion, illumination changes, deformation, and rapid motion. Solving these problems has also become the main task of visual target tracking researchers. As with the development for deep neural networks and attention mechanisms, object-tracking methods with deep learning show great research potential. This paper analyzes the abovementioned difficult factors, uses the tracking framework based on deep learning, and combines the attention mechanism model to accurately model the target, aiming to improve tracking algorithm. In this work, twin network tracking strategy with dual self-attention is designed. A dual self-attention mechanism is used to enhance feature representation of the target from the standpoint of space and channel, with the goal of addressing target deformation and other problems. In addition, adaptive weights and residual connections are used to enable adaptive attention feature selection. A Siamese tracking network is used in conjunction with the proposed dual self-attention technique. Massive experimental results show our proposed method improves tracking performance, and tracking strategy achieves an excellent tracking effect.

1. Introduction

A person's vision serves as a window and a bridge to the world around them. Using it, we have access to a plethora of external sites and the most relevant data available. It is precisely because of the vast amount of information that vision provides us that we can quickly learn and recognize things, and even if they are similar, we can make correct judgments based on our own experience. We acquire up to 80% of the information we receive from the outside world through our vision, demonstrating the critical role it plays in our daily lives. Because visual cognition is so crucial, we need to find a solution to the difficult and grueling problem as soon as possible, even though we all know this. As science and technology have advanced, so has computer vision, which mimics human vision, and different research projects and applications have evolved. In computer vision, image processing, neurobiology, signal processing, machine learning, computers, and imaging technologies are all intertwined. Pattern recognition and mathematics are also important aspects of computer vision. All kinds of objects and important information may be recognized by computers using vision, which is the ultimate evolution of computer vision. During the last few years, computer vision has become one of the most talked about subjects in the world. There's a new era in computer vision thanks to the tremendous advancements in computer science and engineering as well as pattern recognition, image processing, and signal processing as well as neurobiology and machine learning and artificial intelligence. Research in the field of computer vision is focusing on visual object tracking, which uses the newest technology from all three of the abovementioned fields. In intelligent video surveillance, robot autonomous navigation, intelligent traffic management, medical diagnostics, and other sectors, visual target tracking has been widely used [1–5].

Object tracking has gained a lot of attention as a crucial topic in the field of machine vision. It is the process of processing a series of photographs in general. The purpose of moving target tracking is to dig out the predefined feature regions of interest in a series of images in the video and determine their positions, and then associate the feature targets in all images one by one. Target tracking provides reliable data information for object motion behavior and scene analysis and provides powerful help for the correct detection and recognition of moving targets. Therefore, the research and application of target tracking is of great significance. At present, it has been widely used in many fields of production and life, and the future application prospect is even broader. Obviously, the accurate tracking or precise positioning of the target has an immeasurable impact on national security, social development, and people's life stability [6–10].

Object tracking is a hot topic in vision research, with a lengthy history of research and modest progress. However, in recent years, the rapid advancement of science and technology, the breakthrough of big data processing, and the advancement of target tracking have not been as impressive as they once were. Nevertheless, target tracking is still a difficult subject. First, the target scene is generally more complex and changeable, and the object and background to be tracked may even be reversed in different sequences. Second, the uncertainty of the target, the target itself will undergo unpredictable changes, such as pose transformation, defects, blur, and occlusion. Finally, various external and human elements, such as camera shake, weather, lighting, shadows, and other natural conditions, have an impact on tracking. Target tracking is a complicated topic that necessitates long-term research [11–15].

The paper's organization paragraph is as follows: The related work is presented in Section 2. Section 3 analyzes the materials and methods of the proposed work. Section 4, discusses the experiments and results. Finally, in Section 5, the research work is concluded.

2. Related Work

Reference [16] introduced correlation filtering into target tracking, and proposed a novel correlation filter, namely, the minimum output error sum of squares correlation filter. It produced stable correlation filtering during tracking, and the tracker tracked at least 20 times faster than popular trackers at the time. Reference [17] proposed a method of kernel correlation filtering based on MOSSE. In order to solve the problem of lack of samples and redundancy, the method uses circulant matrix to train the sample data. The circulant matrix is then diagonalized by discrete Fourier transform, and the final tracking is achieved by kernel regression. In order to alleviate the edge effect problem in the modified algorithm, the reference [18] proposed a spatial regularization correlation-filtering algorithm, which introduced a penalty term in the loss function to suppress the influence of features far from the target center. Then, the correlation filter is solved by Gauss–Seld iterative method, and the tracking result is finally obtained. Since the algorithm needs to use multiple training images in the online update template stage, it not only increases the time complexity but also makes the performance improvement more difficult. Reference [19] proposes a time-space regularized correlation-filtering algorithm. Based on the SRDCF algorithm, temporal regularization is introduced to a single image, which can be regarded as an approximation of the SRDCF filtering algorithm updating the template through multiple images. Not only the performance and real-time performance are better than the SRDCF algorithm but also the SRDCF algorithm does not have the disadvantage that the tracking is prone to drift when the target apparent change is large. The literature [20] circularly offsets the entire huge image to overcome the drawback that the negative samples of the standard correlation filtering technique are produced by the target image block offset, and the negative samples lack background information outside the target frame. Then, the image block at the target position after the cyclic offset is cropped out as a negative sample to obtain a correlation filter with stronger performance.

Reference [21] proposes visual tracking based on deep learning. The stack auto-encoder is trained offline through a large number of auxiliary images to extract the general features of the image. In the tracking stage, the encoding part of the stack auto-encoder is regarded as a feature extractor. Add a classifier to the last layer of the network and fine-tune the network to accommodate changes in the appearance of objects before tracking. Reference [22] uses a pretrained network to simultaneously extract hierarchical convolutional features, and then learns correlation filters based on these convolutional features to encode the target appearance. Finally, the maximum response value of each layer is calculated to determine the target position. Reference [23] proposed a tracking algorithm based on a deep regression network, using an image set with rectangular box labels to train a neural network offline. In the testing phase, there is no need to fine-tune the network online, and the tracking is performed directly to achieve real-time tracking. Reference [24] proposes a novel multidomain convolutional neural network, which consists of multiple branches of shared layers and domain-specific layers. A model is pretrained by a large number of auxiliary images, and then the shared layer of the pretrained network is combined with the classification layer updated in real time, and finally the tracking is realized. Reference [25] proposed a high-performance visual tracking algorithm based on residual attention Siamese neural network, which relearns correlation filters within the framework of Siamese tracking algorithm. Different kinds of attention mechanisms are introduced to adapt the model without updating the model online, so as to achieve robust tracking. Reference [26] proposes video-based deep reinforcement learning for visual tracking, which consists of three parts. The convolutional neural network extracts image features, the recurrent neural network constructs the video time information and makes full use of the context information between frames, and the agent trained by reinforcement learning is used to make decisions on the position of the target. Reference [27] proposed a deep reinforcement learning tracking based on template matching method, which sent the search image and the apparent template to the shared convolutional layer, and then sent it to the fully connected layer to output the prediction map. The prediction map outputs normalized values under the action of the policy network, and the prediction map corresponding to the largest scalar value is the template closest to the tracking target. Reference [28] proposed a deep reinforcement learning tracking method based on an action decision network, which trains the network through supervised learning in the offline phase to predict the output action in a given state. In the current stage, the neural network is updated through reinforcement learning, and then the network is adaptively fine-tuned in the tracking stage to output 11 possible actions to gradually achieve accurate tracking of the target.

3. Method

This study combines the dual self-attention module with the current popular Siamese network tracking framework and proposes a dual self-attention based Siamese network tracking algorithm. In this study, a dual self-attention module is proposed to solve the problem of insufficient utilization of the features of target template samples and search region samples in Siamese network. It performs self-attention feature enhancement on template features and search features from two aspects of spatial autocorrelation and channel autocorrelation, and adaptively enhances the part of the feature that belongs to the target. This enables the established appearance model to better distinguish the difference between the target and the background, improving the tracking accuracy. The spatial and channel self-attention modules of the dual self-attention module are designed to extract the effective sections of characteristics from two dimensions, respectively, in the dual self-attention module. The adaptive weight is used to fuse the attention features extracted from the two to achieve more efficient self-attention feature extraction. Finally, a target-tracking algorithm based on dual self-attention is designed on the basis of the Siamese tracking network framework. The proposed dual self-attention module is adopted in the multilevel template feature and search feature extraction part to build a more discriminative target appearance model.

3.1. Convolutional Neural Network. The convolutional neural network is extended from the traditional neural network in the spatial dimension and is suitable for data with a two-dimensional spatial structure such as images or videos. It is mainly composed of convolution layer, batch normalization layer, activation layer, pooling layer, fully connected layer, and other structures. Because of their functions, the convolutional layer, batch normalisation layer, and activation layer are closely related, and the network structure formed by the three end-to-end connections is sometimes referred to as a convolutional unit in a convolutional neural network. Convolutional networks often contain such multiple convolutional unit modules, which are stacked on each other to form a complete convolutional neural network model. Convolutional neural networks mainly complete specific tasks by extracting the features of the input image and gradually abstracting them. Convolutional neural network uses multilayer convolution units to extract features from input images, and abstract and pool the extracted features to obtain more concise and highly abstract features. Finally, it is translated into the required results by structures such as fully connected layers designed for specific tasks, and the prediction from images to abstract results is realized.

The convolutional layer is the core structure in the convolutional neural network, and the image data can be regarded as a matrix in a two-dimensional space. The convolution layer performs a convolution operation with each point of the image data and its adjacent points in space through a convolution kernel, obtains the abstract relationship between the point and the surrounding points in space, and generates a corresponding feature map. The computation of discrete convolution is

$$y(k) = h(k) * u(k). \tag{1}$$

The essence of convolution calculation is to transform the input data through the convolution kernel function. Each weight value of the convolution kernel is multiplied by the value at the corresponding spatial position of the input data, and then these multiplication results are summed, and the sum value is the result of the convolution operation. The whole process is similar to the calculation of the convolution kernel sliding on the input data. Compared with the neuron calculation of the traditional neural network, the convolution calculation only considers the adjacent node data, and the parameters of the convolution kernel are shared throughout the sliding calculation process. This mechanism of parameter sharing greatly reduces the amount of parameters in convolutional neural networks.

Compared with shallow neural networks, batch normalization layers are often used in deep neural networks to normalize data to reduce the impact of different data distributions. In the training of deep neural network, due to the excessive number of network layers, the problems of gradient disappearance, and gradient explosion are prone to occur. The root cause is that after the input data is calculated by the activation function for many times, the calculation result will gradually tend to zero value. This causes the calculated gradient values to be biased toward zero or infinity, making it difficult for network training to converge and deviating from the preset training goals. The batch normalization layer can solve the above problems to a certain extent. In the training process of the neural network, in order to make the neural network better, learn the distribution of the entire dataset, multiple data, and corresponding labels are generally packaged and sent to the network for training. Unlike other layers that perform computations directly on the input data, the batch normalization layer operates on the dimension of batches of data. Batch normalization is calculated as follows:

$$x' = \frac{(x-\mu)}{\sqrt{\sigma^2+\varepsilon}},$$
$$y = \alpha x' + \beta. \tag{2}$$

After batch normalization of each batch of data, the data can be remapped to the normalized space and the convergence speed of the network can be accelerated. In addition to suppressing the problem of extremely large or small gradients to a certain extent, it also enables the neural network to better learn the ability to extract data content. This avoids the network's preference for some data with a high proportion in the dataset and reduces the overfitting of the model to a certain extent. In addition, the linear mapping parameters of the batch normalization layer can avoid that the training data are always located in a single normalized space, which improves the representation ability of the network.

In the convolutional neural network structure, the activation layer is a key component required for the network model to be able to fit nonlinear functions. Since the convolution computation is a linear mapping, it is difficult for the nesting of linear functions to learn the implicit nonlinear relationship from data to labels. Therefore, it is necessary to add nonlinear mapping to the network model, so that the network has the ability to learn nonlinear relationships. The activation layer can use a nonlinear function to truncate or compress the input data into a limited space and retain the effective part of the output features of the previous layer. This not only enables the network to complete more abstract and complex tasks but also inhibits the spread of information generated by invalid nodes, making network learning more directional.

Convolutional neural networks that can learn abstract features use pooling layers as a major component. By replicating the human visual system, the pooling layer decreases the dimension of the input data, picks more representative characteristics to represent the image content, and accomplishes the function of extracting the image's abstract information. In addition, the pooling layer reduces the scale of features and improves the training efficiency of the network while filtering the significantly abstract information in the input data.

The fully connected layer is the structure of the traditional neural network, and it has also been widely used in the convolutional neural network. The fully connected layer is generally set at the last layer of the network and maps the abstract features extracted by the previous network modules to the required tasks. Taking the classification task as an example, the number of output nodes of the fully connected layer is the number of all categories. In addition, the fully connected layer also has the function of feature compression and recovery. By setting the number of upper- and lower-layer nodes with different numbers, the input features can be encoded and compressed or decoded and recovered.

3.2. Siamese Network with Dual Self-Attention. Figure 1 depicts the dual self-attention Siamese network tracking (DSASN) system presented in this article. The ResNet-50 backbone network is first utilized to extract the relevant multilevel convolutional features (MLCF) from the target template picture and the search region image. This feature consists of the outputs of three layers of convolutional units at different depths in the backbone network. Then the template feature adjustment layer and the search feature adjustment layer are used to adjust the two, respectively, and the adjusted multilevel template features and multilevel search features are obtained. Then, the respective dual self-attention (DSA) modules are constructed for the template branch and the search branch, respectively, and the dual self-attention enhancement is performed on each layer feature in the multilevel features of the two. Then, the classification and regression network is used to calculate the template features and search features of each layer enhanced by double attention, and the classification and regression feature maps corresponding to the features of each layer are obtained. Finally, the classification and regression results are achieved by adaptively fusing the weights of the classification and regression feature maps calculated by the three classifications and regression networks. The adaptive fusing of multilevel feature prediction findings is used in the Siamese network framework. By training the respective adaptive weights for the classification and regression network corresponding to each layer of features, the influence of features at different levels of abstraction is fully considered.

The structure of the classification and regression network is shown in Figure 2. For the classification and regression tasks, the classification and regression networks create matching deep cross-correlation layers. Each task uses its own similarity feature for subsequent calculation, which avoids the problem of task preference caused by similarity feature sharing between two tasks.

3.3. Dual Self-Attention Module. The target itself changes during the tracking process and is easily disturbed by similar or cluttered backgrounds, resulting in tracking errors by the tracker. How to accurately model the target and improve the ability of the network to discriminate between the target and the background is the key to improving the performance of the tracker. Therefore, this chapter proposes a dual self-attention mechanism module to enhance the part of deep features belonging to the target from two dimensions of space and channel and improve the ability of the Siamese network to distinguish the target from the background.

The main structure of the spatial self-attention module (SSA) is shown in Figure 3. The spatial self-attention module enhances the part of the input features belonging to the target region from the perspective of spatial location. The spatial self-attention module calculates the correlation between each vector of the input feature in the spatial dimension and other vectors to generate the spatial significance matrix of the spatial feature vector. The original features are then spatially weighted with the normalized spatial significance matrix. Finally, residual connection is performed with the input feature to obtain the salient feature representation of the feature in the spatial dimension.

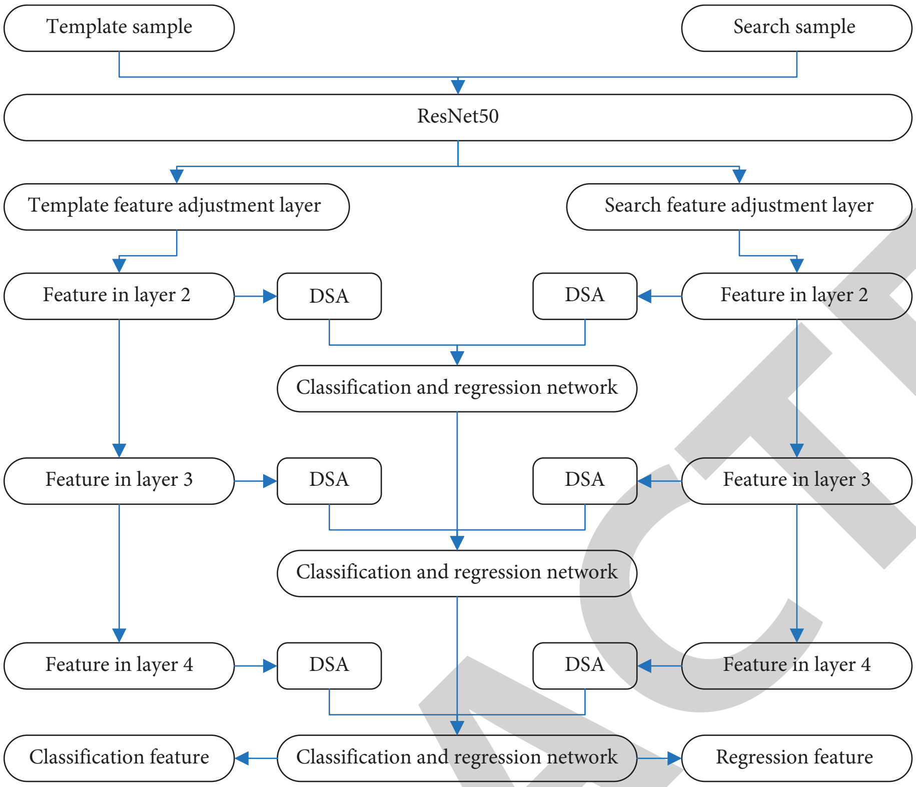

Figure 1: Architecture of dual self-attention based Siamese tracking network.

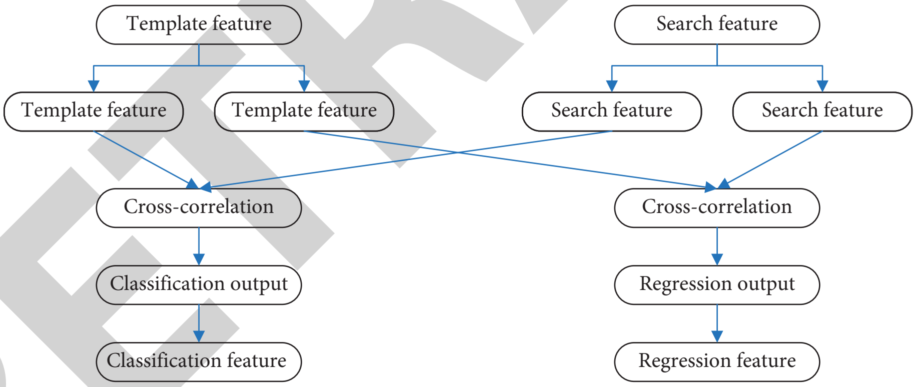

Figure 2: Architecture of classification and regression network.

The spatial self-attention module first performs Query, Key, and Value transformations on the input features to obtain the channel-compressed input features. Different from the nonlocality module, the three transforms employed in the spatial self-attention module are bottleneck layers with kernel size 1. The number of output channels is smaller than the number of input channels in order to reduce the weight of the subsequent calculation of the spatial importance matrix. Since the spatial self-attention is mainly aimed at the feature vectors in the spatial dimension, the calculation amount can be reduced by shrinking the channel dimension, and the participation of invalid channels in the subsequent calculation of the spatial importance matrix can also be reduced. Then, the compressed features obtained by Query transformation and Key transformation are rearranged. The spatial importance matrix is the following:

$$m_{ji} = \exp \frac{\left(q_i k_j\right)}{\sum_{i=1}^{N} \exp\left(q_i k_j\right)}. \tag{3}$$

After getting the spatial importance matrix, rearrange the value transformed features and multiply it by the spatial importance matrix. The obtained results are rearranged and added to the input features to obtain spatial self-attention features. The spatial self-attention feature is the following:

$$E_j = \alpha \sum_{i=1}^{N} \left(m_{ji} v_i\right) + x_j. \tag{4}$$

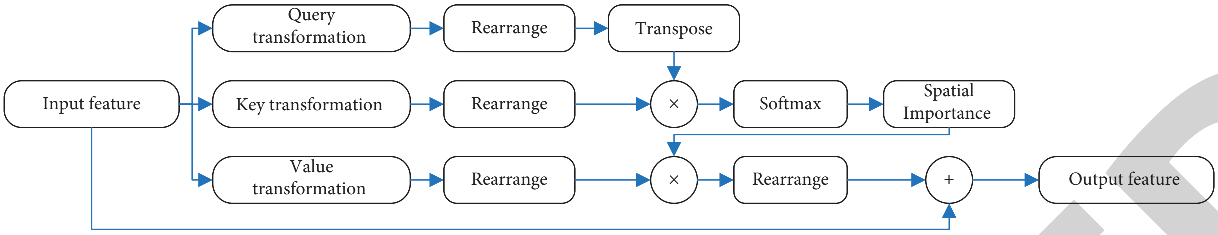

Figure 3: Architecture of spatial self-attention module.

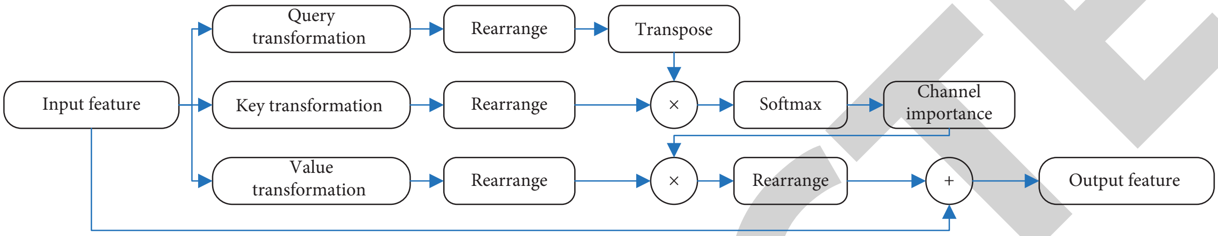

Figure 4: Architecture of channel self-attention module.

Each spatial vector on the spatial self-attention feature is obtained by the sum of the similarity between the spatial vector and all the spatial vectors. Therefore, the spatial self-attention module realizes the self-attention enhancement to the spatial dimension of the input feature and strengthens the part belonging to the target from the perspective of spatial position.

The main structure of the channel self-attention module (CSA) is shown in Figure 4. The channel self-attention module extracts the part of the input features belonging to the target category from the category perspective. The channel self-attention module obtains the importance matrix of the channel planes by calculating the correlation between each channel plane and other channel planes in the channel dimension of the input features. The normalized channel importance matrix is then channel-weighted with the original features. Finally, residual connection is performed with the input feature to obtain the salient feature representation of the feature in the channel dimension.

Different from the spatial self-attention, the channel self-attention module obtains the Query, Key, and Value features by rearranging the input features. It does not use transformation to reduce the number of channels, and retains more channel information for subsequent self-attention channel selection. The channel self-attention is calculated as follows:

$$n_{ji} = \exp \frac{\left(q_i k_j\right)}{\sum_{i=1}^{C} \exp\left(q_i k_j\right)}. \tag{5}$$

After getting the channel importance matrix, it is matrix-multiplied with the value feature. The obtained results are rearranged and added to the input features point by point, and finally the channel self-attention feature is obtained:

$$F_j = \beta \sum_{i=1}^{C} \left(n_{ji} v_i\right) + x_j. \tag{6}$$

Each channel plane on the channel self-attention feature is obtained by the sum of the similarity between the channel plane and all channel planes. Therefore, the channel self-attention module realizes the self-attention enhancement of the input feature channel dimension and strengthens the part that more accurately represents the target category from the perspective of channel category. After obtaining the spatial self-attention feature and the channel self-attention feature of the input feature, the dual self-attention module adopts an adaptive fusion method to weight and fuse the two to obtain the dual self-attention feature. The calculation of the double self-attention feature is the following:

$$G = \mathrm{AE} + \mathrm{BF}, \tag{7}$$

A and *B* are adaptively learned during the training process of the network to achieve more efficient spatial channel self-attention feature fusion.

3.4. Loss Function.

To teach the dual self-attention Siamese network how to forecast the target's location and size at the same time, this chapter adopts the multitask learning loss function of joint classification and regression to train the dual self-attention siamese network. The loss function is the following:

$$L = \lambda_1 L_{\mathrm{cls}} + \lambda_2 L_{\mathrm{reg}}. \tag{8}$$

A multitask learning loss function is created by combining the classification and regression losses and setting their weights to be equal. This enables the network to give equal attention to the classification and regression tasks during the learning process and improves the accuracy of the network's dependence on the classification prediction results during the tracking process.

4. Experiment and Discussion

In this section, we define the dataset and metric, method comparison, evaluation on attention, evaluation on

Table 1: Experiment environment information.

Item	Detail
Operating system	Ubuntu 18.04
CPU	i7-6700
Memory	64 GB
Deep framework	PyTorch 1.7

attention, Evaluation on Feature Fusion, and evaluation on loss in detail.

4.1. Dataset and Metric. The dataset used in this work is the VOT2019 dataset, which is a public dataset used in the 2019 Visual Object Tracking Challenge. It contains 60 video sequences, which is different from 12 video sequences in VOT2018. The evaluation indicators of the VOT2019 dataset used in this work are EAO and Accuracy. The experimental platform information is illustrated in Table 1.

4.2. Method Comparison. To verify the feasibility of DSASN designed in this work, it is first compared with other visual target tracking algorithms, including SiamRPN++ [29], SiamMask [30] and SiamCRF_RT [31]. The experimental results are illustrated in Table 2.

The method proposed in this work achieves the highest performance. Compared with the best-performing SiamRPN++ method in the table, 2.4% EOA improvement and 2.2% ACC improvement can be obtained, respectively. This verifies the advancement and feasibility of the DSASN method.

4.3. Evaluation on Attention. This work uses a dual self-attention mechanism to improve the feature learning ability of the network. This paper conducts comparative studies to examine target tracking performance without DSA and when DSA is applied in order to verify the effectiveness of this method. The experimental results are illustrated in Figure 5.

Compared with not using DSA, after using DSA, we can get 3.2% increase in EOA and 5.9% increase in ACC. It can be proved that the attention mechanism used in this work can effectively improve the target tracking performance.

Going a step further, the DSA attention mechanism consists of TSA and CSA. In order to verify that combining these two different mechanisms can maximize network performance, this work conducts comparative experiments to compare the performance of using a single TSA and a single CSA, respectively. The experimental results are illustrated in Figure 6.

It can be seen from the figure that using a single-attention mechanism does not achieve the best-tracking performance. Network performance can only be maximized by combining the two different attention methods.

Table 2: Result of method comparison.

Method	EAO	ACC
SiamRPN++	28.5	59.9
SiamMask	28.3	59.4
SiamCRF_RT	26.2	54.9
DSASN	30.9	62.1

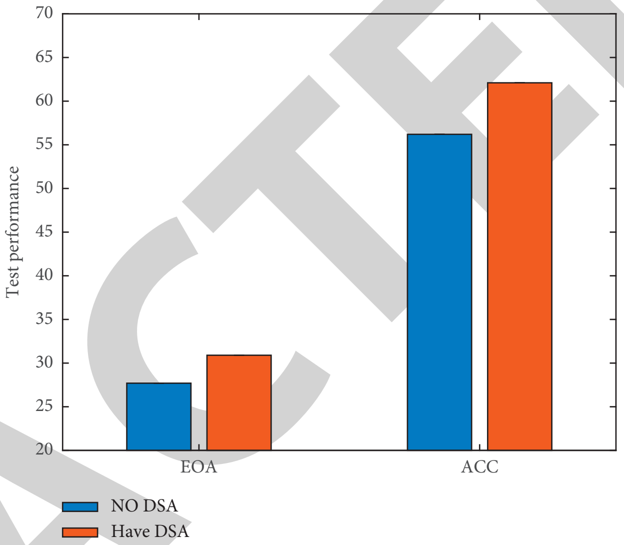

Figure 5: Comparison of without DSA and with DSA.

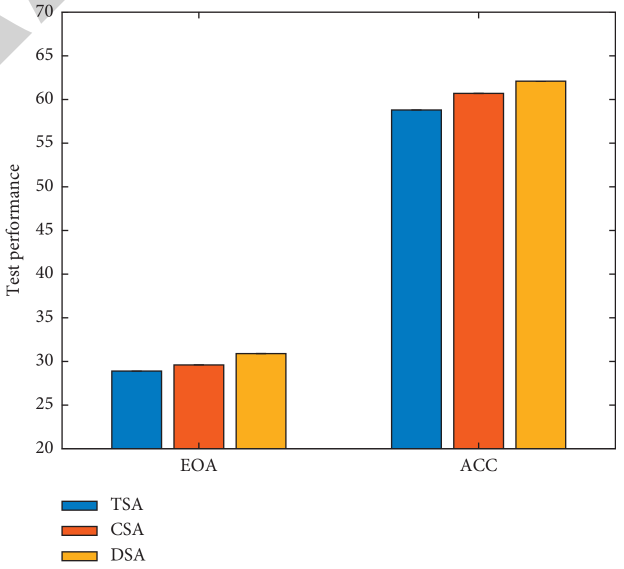

Figure 6: Comparison of different attention.

4.4. Evaluation on Feature Fusion. As mentioned earlier, this work fuses features from different levels. To verify the effectiveness of this fusion strategy, this work compares the tracking performance when using single-layer features and multilayer features, respectively. The experimental results are illustrated in Figure 7.

Compared with not using multiple feature, after using it, we can get 1.6% increase in EOA and 2.9% increase in ACC. It can be proved that the multiple feature strategy used in this work can effectively improve the target tracking performance.

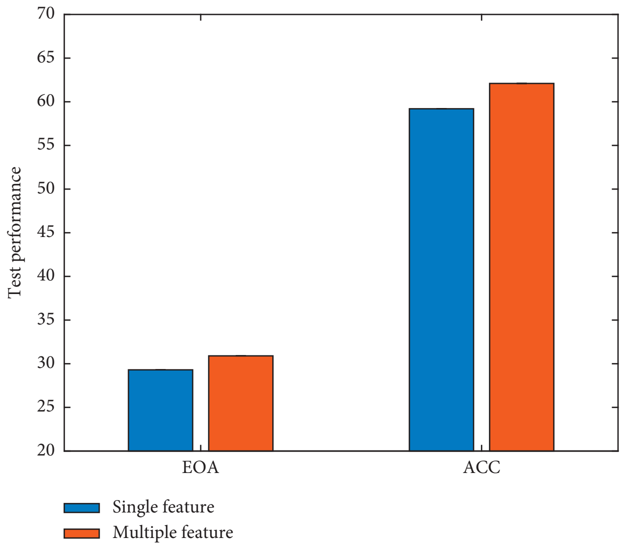

Figure 7: Evaluation on feature fusion.

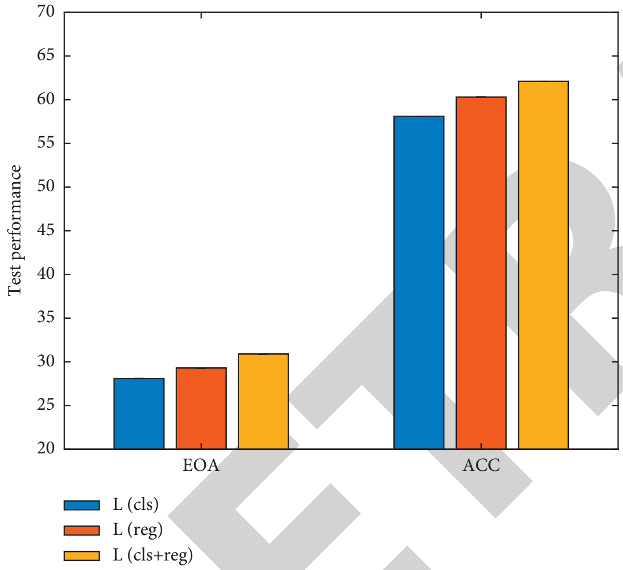

Figure 8: Evaluation on different losses.

4.5. Evaluation on Loss. This paper proposes a loss that is a combination of classification and regression loss. This work conducts comparison tests to assess the tracking performance while employing a single loss and a combination loss, respectively, to verify the usefulness of this technique. The experimental results are illustrated in Figure 8.

It can be seen from the figure that using a single-loss function does not achieve the best-tracking performance. Network performance can only be maximized by combining the two different losses.

5. Conclusion

Target tracking research and application has a significant impact on computer vision progress. Target tracking is a difficult study direction in computer vision because it is such an essential field. Although many scholars have worked hard for decades, the effect of target tracking has also been significantly improved, but there is no tracking algorithm that can handle various complex scenarios well. The emergence of deep learning methods makes it possible to build more robust-tracking methods. In this paper, we design a new target feature representation combined with the attention mechanism, aiming to improve the performance of the tracking algorithm. This work proposes a twin network-tracking algorithm based on dual self-attention. A dual self-attention module is built by examining the problem of insufficient utilization of target templates and search features in current Siamese network-tracking techniques. To accurately simulate the target appearance, it mines more effective components from the target feature information. The dual self-attention module enhances the self-attention of the part belonging to the target from the two dimensions of space and channel, and enhances the ability of the target appearance model to distinguish the target from the background. The dual self-attention module is made up of two modules: one for spatial self-attention and another for channel self-attention, which calculates the self-attention weights of features in two dimensions of space and channel, respectively. Adaptive weights execute the fusion, allowing the model to automatically select more effective feature dimensions. The algorithm proposed in this work uses dual self-attention modules to enhance the template features and search features in the twin-tracking network, respectively, and improves the discriminability of the tracking network by increasing the weight of the target part in the feature. Experimental results demonstrate that the proposed dual self-attention module significantly improves the tracking performance.

Data Availability

The datasets used during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Xu, J. Wang, W. Shou, T. Ngo, A.-M. Sadick, and X. Wang, "Computer vision techniques in construction: a critical review," *Archives of Computational Methods in Engineering*, vol. 28, no. 5, pp. 3383–3397, 2021.

[2] C. Z. Dong and F. N. Catbas, "A review of computer vision-based structural health monitoring at local and global levels," *Structural Health Monitoring*, vol. 20, no. 2, pp. 692–743, 2021.

[3] H. Tian, T. Wang, Y. Liu, X. Qiao, and Y. Li, "Computer vision technology in agricultural automation -A review," *Information Processing in Agriculture*, vol. 7, no. 1, pp. 1–19, 2020.

[4] K. Chandra, A. S. Marcano, S. Mumtaz, R. V. Prasad, and H. L. Christiansen, "Unveiling capacity gains in ultradense networks: using mm-wave NOMA," *IEEE Vehicular Technology Magazine*, vol. 13, no. 2, pp. 75–83, 2018.

[5] J. Guo, H. He, T. He et al., “GluonCV and GluonNLP: deep learning in computer vision and natural language processing,” *Journal of Machine Learning Research*, vol. 21, no. 23, pp. 1–7, 2020.

[6] S. Kamkar, F. Ghezloo, H. A. Moghaddam, A. Borji, and R. Lashgari, “Multiple-target tracking in human and machine vision,” *PLoS Computational Biology*, vol. 16, no. 4, Article ID e1007698, 2020.

[7] J. Du, C. Jiang, and Z. Han, “Contract mechanism and performance analysis for data transaction in mobile social networks,” *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 2, pp. 103–115, 2017.

[8] P. Dendorfer, A. Osep, A. Milan et al., “MOTChallenge: a benchmark for single-camera multiple target tracking,” *International Journal of Computer Vision*, vol. 129, no. 4, pp. 845–881, 2021.

[9] A. Altan and R. Hacıoğlu, “Model predictive control of three-axis gimbal system mounted on UAV for real-time target tracking under external disturbances,” *Mechanical Systems and Signal Processing*, vol. 138, Article ID 106548, 2020.

[10] J. Yan, W. Pu, S. Zhou, H. Liu, and Z. Bao, “Collaborative detection and power allocation framework for target tracking in multiple radar system,” *Information Fusion*, vol. 55, pp. 173–183, 2020.

[11] S. S. Moghaddasi and N. Faraji, “A hybrid algorithm based on particle filter and genetic algorithm for target tracking,” *Expert Systems with Applications*, vol. 147, Article ID 113188, 2020.

[12] W. Yi, Z. Fang, W. Li, R. Hoseinnezhad, and L. Kong, “Multi-frame track-before-detect algorithm for maneuvering target tracking,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4104–4118, 2020.

[13] C. Xu, X. Wang, S. Duan, and J. Wan, “Spatial-temporal constrained particle filter for cooperative target tracking,” *Journal of Network and Computer Applications*, vol. 176, Article ID 102913, 2021.

[14] J. Liu, Z. Wang, and M. Xu, “DeepMTT: a deep learning maneuvering target-tracking algorithm based on bidirectional LSTM network,” *Information Fusion*, vol. 53, pp. 289–304, 2020.

[15] Y. Wang, T. Wang, G. Zhang, Q. Cheng, and J.-q. Wu, “Small target tracking in satellite videos using background compensation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7010–7021, 2020.

[16] V. H. Diaz-Ramirez, V. Contreras, V. Kober, and K. Picos, “Real-time tracking of multiple objects using adaptive correlation filters with complex constraints,” *Optics Communications*, vol. 309, pp. 265–278, 2013.

[17] J. F. Henriques, R. Caseiro, and P. Martins, “High-speed tracking with kernelized correlation filters,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

[18] D. Zhang, Z. Zhang, L. Zou et al., “Part-based visual tracking with spatially regularized correlation filters,” *The Visual Computer*, vol. 36, no. 3, pp. 509–527, 2020.

[19] S. Feng, K. Hu, E. Fan, L. Zhao, and C. Wu, “Kalman filter for spatial-temporal regularized correlation filters,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3263–3278, 2021.

[20] X. Sheng, Y. Liu, H. Liang, F. Li, and Y. Man, “Robust visual tracking via an improved background aware correlation filter,” *IEEE Access*, vol. 7, pp. 24877–24888, 2019.

[21] J. Kuen, K. M. Lim, and C. P. Lee, “Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle,” *Pattern Recognition*, vol. 48, no. 10, pp. 2964–2982, 2015.

[22] C. Ma, J. B. Huang, and X. Yang, “Robust visual tracking via hierarchical convolutional features,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2709–2723, 2018.

[23] C. Fang, J. Huang, K. Cuan, X. Zhuang, and T. Zhang, “Comparative study on poultry target tracking algorithms based on a deep regression network,” *Biosystems Engineering*, vol. 190, pp. 176–183, 2020.

[24] J. Zhang, K. Zhao, and B. Dong, “Multi-domain collaborative feature representation for robust visual object tracking,” *The Visual Computer*, vol. 37, no. 9, pp. 2671–2683, 2021.

[25] X. Sun, G. Han, L. Guo, H. Yang, X. Wu, and Q. Li, “Two-stage aware attentional Siamese network for visual tracking,” *Pattern Recognition*, vol. 124, Article ID 108502, 2022.

[26] D. Zhang, H. Maei, and X. Wang, “Deep reinforcement learning for visual object tracking in videos,” 2017, http://arXiv.org/abs/1701.08936.

[27] J. Choi, J. Kwon, and K. M. Lee, “Visual tracking by reinforced decision making,” p. 2, 2017, http://arXiv.org/abs/1702.06291.

[28] Y. Jiang, D. K. Han, and H. Ko, “Relay dueling network for visual tracking with broad field-of-view,” *IET Computer Vision*, vol. 13, no. 7, pp. 615–622, 2019.

[29] B. Li, W. Wu, and Q. Wang, “Siamrpn++: evolution of siamese visual tracking with very deep networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4282–4291, Long Beach, CA, USA, June 2019.

[30] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, “Fast online object tracking and segmentation: a unifying approach,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1328–1338, Long Beach, CA, USA, 2019.

[31] G. Wang, C. Luo, and Z. Xiong, “Spm-tracker: series-parallel matching for real-time visual object tracking,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3643–3652, Long Beach, CA, USA, June 2019.