Computers, Materials & Continua
DOI: 10.32604/cmc.2022.028896
Article

Tech Science Press

Criss-Cross Attentional Siamese Networks for Object Tracking

Zhangdong Wang[1], Jiaohua Qin[1,*], Xuyu Xiang[1], Yun Tan[1] and Neal N. Xiong[2]

[1]College of Computer Science and Information Technology, Central South University of Forestry & Technology, Changsha, 410004, China
[2]Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, 74464, OK, USA
*Corresponding Author: Jiaohua Qin. Email: qinjiaohua@csuft.edu.cn
Received: 20 February 2022; Accepted: 10 April 2022

Abstract: Visual object tracking is a hot topic in recent years. In the meanwhile, Siamese networks have attracted extensive attention in this field because of its balanced precision and speed. However, most of the Siamese network methods can only distinguish foreground from the non-semantic background. The fine-tuning and retraining of fully-convolutional Siamese networks for object tracking(SiamFC) can achieve higher precision under interferences, but the tracking accuracy is still not ideal, especially in the environment with more target interferences, dim light, and shadows. In this paper, we propose criss-cross attentional Siamese networks for object tracking (SiamCC). To solve the imbalance between foreground and non-semantic background, we use the feature enhancement module of criss-cross attention to greatly improve the accuracy of video object tracking in dim light and shadow environments. Experimental results show that the maximum running speed of SiamCC in the object tracking benchmark dataset is 90 frames/second. In terms of detection accuracy, the accuracy of shadow sequences is greatly improved, especially the accuracy score of sequence HUMAN8 is improved from 0.09 to 0.89 compared with the original SiamFC, and the success rate score is improved from 0.07 to 0.55.

Keywords: Criss-cross attention; object-tracking; siamese-network

1 Introduction

Visual object tracking is a basic problem in the fields of computer vision analysis [1,2], automatic driving, and video supervision. It needs to track the object automatically in the changing video sequence. The key problem of target tracking is how to detect and locate the target quickly and accurately under the conditions of occlusion, shadows, out of sight, deformation, and background clutter.

Modern trackers can be roughly divided into two branches [3,4]. The first branch is object tracking based on a correlation filter, which operates in the Fourier domain by the property of cyclic correlation to train the regressors. The weights of filters can be updated effectively in real-time while tracking online. The method based on a correlation filter improves the accuracy of the model through deep

features, but it has a great influence on the speed of updating the model [5,6]. The second branch is object tracking based on deep learning [7,8]. It pre-trains the target directly by using a deep convolutional neural network (CNN), which has been demonstrated to be very successful in the visual field recently [9,10], and then it optimizes the deep network by using stochastic gradient descent (SGD) [11]. However, the limited training data and a large number of parameters make it difficult to pre-train the network, while SGD requires a longer run time to achieve high-precision tracking.

In recent years, the Siamese convolution network based on deep learning has attracted great attention in the field of visual tracking because of its good performance. The Siamese network tracker transforms the problem of visual object tracking into learning the cross-correlation between the features of the target template and the features of the search area through offline training, to find the similarity mapping of their universality. Tao et al. [12] innovatively transforms the problem of target tracking into the problem of patch block matching. Bertinetto et al. [13] designs an end-to-end network and proposes a higher level of siamese similarity function, which greatly improves the tracking speed. Wang et al. [14] introduces an attention module into object tracking, in which multiple attention mechanisms take into account the temporal and spatial information of video, and the over-fitting of the network is effectively alleviated. At the same time, the discriminant ability of the network is improved. Valmadre et al. [15] transforms the correlation filtering operation into a single network-layer embedded in SiamFC. Guo et al. [16] embeds the target appearance variation and background suppression into SiamFC, which enhances the online updating ability of the model. Li et al. [17] introduces region proposal network after Siamese network(Siam RPN), which describes tracking as a one-time local detection task, and improves the speed and accuracy at the same time. Zhu et al. [18] further introduces a distractor-aware module based on Siam RPN, and improves the discrimination power of the model. Fan et al. [19] uses a cascade architecture to deal with the non-uniformity of sample blocks, and uses multiple steps of regressions and feature transform blocks to obtain more accurate information. Wang et al. [20] adds mask branch based on SiamFC, and solves the problem of segmentation and tracking. Li et al. [21] improves the original sampling strategy and uses a multi-cascade combination to improve the robustness of features. At this point, the trackers based on the Siamese network can surpass trackers based on correlation filtering.

Most Siamese network trackers only distinguish foreground and non-semantic background. The semantic background is usually treated with a separator. When the background is messy, dimly lit, and shaded, splitter is not guaranteed, and it is impossible to distinguish target and distractors from a background in a complex environment. Such as Captcha recognition, CNN has good performance to eliminate the background noise by end-to-end learning [22], rather than the manual feature [23]. SiamFC uses a weighted loss function to eliminate the class imbalance of the positive and negative examples, which is ineffective. We repeated SiamFC and analyzed 8 sequences with poor effect in the evaluation of object tracking benchmark. These 8 sequences can be divided into two categories: one is sequence bolt and football, whose poor effect is due to the high similarity between the target and other interferences. Another is HUMAN8 and IRONMAN, whose first target environment is dark, which makes it difficult to select features and track them correctly. For the first kind, we fine-tune SiamFC using the Got10k [24] dataset on top of the original parameters. After fine-tuning and retraining, SiamFC greatly improves the tracking accuracy of BOLT and FOOTBALL sequences. However, the sequence tracking accuracy of the second kind with shadow and illumination poor environment in the first frame still has no improvement.

We find that the low accuracy of video tracking in shadow and illumination poor environment in the first frame is because of the imbalance of the semantic and non-semantic background of the features, which makes the tracker unable to distinguish the target and background interferences.

Semantic segmentation is conducive in many applications, from biometric recognition to target recognition. Since a natural object may produce numerous images with a different appearance, perspective, posture, illumination, and complex background [25], it is quite challenging. Therefore, from the level of semantic segmentation of object segmentation, this paper introduces a criss-cross attention module to improve the accuracy of video tracking.

Most target tracking algorithms only focus on the processing of regional information when extracting image features. These algorithms neglect the correlation between different locations in the image, resulting in incomplete features and loss of context information, which affects the learning of feature weights. Due to the non-locality of the traditional method in CNN convolution operation, the convolution kernel size affects its receptive field, so all convolution kernels get the feature local area information. However, relative information between distant pixels is also valuable, especially when processing object tracking based on video frame data. When tracking the movement of a character. We should pay more attention to the relative positions of arms and legs, analyze from the perspective of the overall moving target, and reduce the influence of background and interference on the target from the perspective of comparison.

The attention model is widely used in various tasks [26], which can enhance the representational power of the network by modeling the channel relations in the attention mechanism [27]. Chen et al. [28] used several attention masks to fuse feature maps or predictions from different branches. Pan et al. [29] used data augment to train the model. Wang et al. [30] proposed a non-local method, which generated a huge attention graph by calculating the relationship matrix between each spatial point in the feature graph, and then paid attention to guide intensive contextual information aggregation. Fu et al. [31] used the self-attention model to obtain contextual information. Zhao et al. [32] learned an attention map to aggregate contextual information for each point adaptively and specifically. Huang et al. [33] put forward the criss-cross module, which collected contextual information in horizontal and vertical directions, processed information globally, and better analyzed the integrity of the image. It strengthened the correlation between the target and non-target, and can add the correlation information between the remote pixels into the object tracking network, to obtain more informative feature maps and capture the full text of the message [34], which can improve the performance of tracking network.

For the complex tracking task and high-quality training data, we propose a criss-cross attention Siamese network for end-to-end offline training based on large-scale image pairs. It includes Siamese sub-network for feature extraction and a cross-attention sub-network for feature enhancement. We use two branches of Siamese sub-network to extract the features of template frames and detection frames simultaneously. Aiming at the imbalance between non-semantic background and semantic interference, we firstly use dilation convolutions on the output feature map of template frame branches of Siamese sub-network to obtain more intensive information. Then we use the cross-attention module to analyze the correlation within the template frame feature map, to distinguish more accurately the semantic, non-semantic background, and semantic background, and improve the overall situation. The cross-correlation function is used to calculate the end output characteristic graph of the two branches to get the score map. Finally, the score map is transformed into the tracking block diagram. SiamCC network effectively solves the tracking problem in the shadow and illumination poor environment of the first frame. In the object tracking benchmark(OTB100) [35], SiamCC runs at a maximum speed of 90 frames per second. In terms of accuracy, the sequence accuracy of dim light and shadow has been greatly improved. Especially, the HUMAN8 accuracy score of the sequence has been improved from 0.09 to 0.89 compared with the original SiamFC, and the success rate has been improved from 0.07 to 0.55.

In this paper, we propose a novel Siamese network object tracking method based on criss-cross attention feature enhancement. The contributions can be summarized as the following three aspects.

1. We evaluate and analyze the SiamFC, and find two main reasons for the poor object tracking: the influence of target similar interferences and the limit of dim light and shadow.
2. We have fine-tuned and retrained the SiamFC. Compared with the results of the original SiamFC, the overall accuracy is slightly improved, which solves the impact of object similar interferences in a small range.
3. We propose a SiamCC network that uses the cross focus module to improve the correlation between features of the whole image, which strengthens the target features and greatly improves the tracking effect of poor light and shadow environment.

The rest of this paper is organized as follows. In Section 2, we give a brief review of two related works on the Fully-Convolutional Siamese network and Criss-Cross network. Section 3 describes the proposed SiamCC approach. Section 4 presents the performance analysis. The conclusions are drawn in Section 5.

2 Related Works

In this section, we briefly review two algorithms related to our work: Fully-Convolutional Siamese Network (SiamFC) and Criss-Cross Network (CCNet).

2.1 Fully-convolutional Siamese Network

Object tracking can be regarded as the learning of similarity problems in the initial offline stage. Siamese network is the typical learning of similarity in a deep convolution network. The Siamese network learns similarity measures from the data, compares and matches them with the samples of new unknown classes, then the object tracking problem is transformed into the matching problem of patch blocks. The Siamese neural network has two-branched neural networks with identical weights of structural parameters. The two neural networks map the different input information to the new space to generate the feature matrix. By calculating the loss function, the similarity between the two inputs is evaluated. Based on the Siamese network, SiamFC uses a cross-correlation function to compare the similarity between the sample image z and search image x. If the similarity calculation scores of two images are high, they are likely to be judged as the same target, and vice versa, the probability of the target is low. The advantage of the fully-convolutional network is that the retrieval frame does not need to have the same size as the template frame. It can provide a larger search image as input for the network, and then calculate the similarity of all translation windows on the dense grid.

2.2 Criss-cross Attention for Semantic Segmentation

Semantic segmentation, object recognition, object tracking, and other fields [36] are all used to process image feature information [37]. If long-distance context-dependent information is captured, the performance of the algorithm can usually be improved at a level. Previously, the principle of capturing long dependencies was to use multiple dilation convolutions or utilize pyramid pooling modules. However, dilation convolution can only collect information from a small number of surrounding pixels, and cannot generate accurate and dense contextual information [38]. The pooling method aggregates contextual information in a non-adaptive manner and the homogeneous contextual information is adopted by all image pixels, but it can't satisfy the requirement of different contextual information for different pixels. In the traditional method of generating dense pixel-level context information,

Zhao et al integrates the context information of each location on the predicted attention map. Non-local Networks utilize a self-attention mechanism, which enables a single feature from any position to be able to perceive the features of all other locations, thereby generating more power pixel-wise representation. Each location in the feature map is connected to all other locations by using a self-adaptively predicted attention map, to obtain contextual information of various ranges. However, the traditional method based on attention mechanism needs to calculate the relationship between each pixel, which will spend huge calculation. Based on the attention mechanism, CCnet uses two consecutive criss-cross attention operations instead of non-local operations to aggregate the contextual information of remote pixels horizontally and vertically.

3 Our Proposed SiamCC System

Through the analysis of tracking results in light and dark environments, we find that non-semantic background, semantic interference, and poor global correlation of features are the reasons for poor tracking results. Therefore, we introduce the repetitive criss-cross attention module in the full-convolution Semantic network, which is used to capture contextual information from remote dependency in a more effective way, effectively distinguishing between semantic background and non-semantic background, highlight target features, and greatly improve the tracking accuracy in the dark environment.

3.1 Overall Network Framework of SiamCC

The network flow of SiamCC is shown in Fig. 1. The input image generates features Z' and X' by template branches and detection branches of the full-convolution Siamese network respectively. To retain more details and generate dense features effectively, we use dilation convolutions for the feature Z' of template frame, which enlarges the width and height of the output feature mapping of template frame, and obtains the feature matrix Z'' with the length and width as $H \times W$. After that, we use the convolutional layer to get the reduced dimension feature map H. Then we feed the feature map H to the vertical and horizontal criss-cross attention module, which computes and analyzes the organizational contextual information for each pixel by cross-over to generate a new feature map H'. However, feature mapping H' only aggregates contextual information in horizontal and vertical directions, which is not strong enough to distinguish semantic background from the non-semantic background. To obtain richer and more intensive contextual information, we provide feature mapping H' to the criss-cross attention module and get feature mapping H''. Therefore, each location in the feature map H'' actually collects information from all the pixels. It should be noted that the recurrent criss-cross attention modules share the same parameters and avoid too many additional parameters. Then we connect the dense contextual feature H'' with the local representation feature Z'', and batch normalization and activation feature fusion of one or more convolutional layers to obtain the global enhancement feature. Finally, the global enhancement feature and the detection frame feature are operated with cross-correlation and scored to get the target area of the next frame.

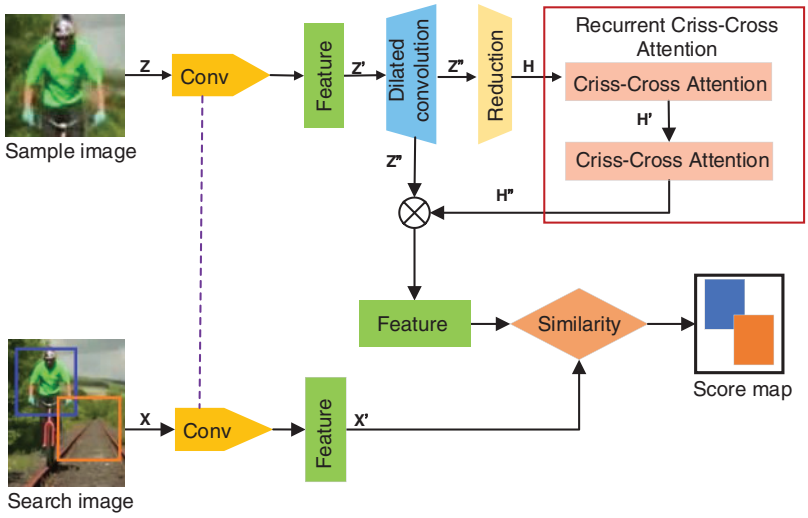

Figure 1: The network architecture of SiamCC

3.2 Feature Enhancement Based on Criss-cross Attention Module

Through the theoretical analysis and experimental demonstration of the SiamFC tracker, we found that the poor tracking effect of SiamFC under poor light and shadow environment is due to the poor global characteristics, which could not effectively separate the target from the interfering substances.

Based on SiamFC feature extraction network, we use the criss-cross attention module to enhance the features, which collects the contextual information in horizontal and vertical directions, and processes the information globally. It can better analyze the integrity of the image and strengthen the analysis of the correlation between the target and non-target. It can also incorporate the correlation information between remote pixels into the object tracking network to obtain more informative feature maps and capture full-text information, then further improve the performance of the tracking network. The basic mechanism of the criss-cross attention module is the attention mechanism. Its concrete manifestation and calculation process is as Eq. (1):

$$y_i = \frac{1}{C(x)} \sum_{\forall_j} f(x_i, x_j)\, g(x_j) \tag{1}$$

where x_j represents each pixel of the input image, f calculates the correlation between x_i and all x_j ,g can be understood as an enhancement of x_j Notice that x is a vector. So the result of f is a numerical value, and $g(x_j)$ is still a vector. The vector x_j at each pixel of the input image is multiplied by the weight value of the correlation between x_i and x_j, and the more the correlation, the greater the contribution to the result. After calculating all the weighted results of x_j, the current y_i is obtained, at which time y_i contains the relevant information between x_i and all the surrounding pixels. The general idea of the self-attention mechanism is the matrix multiplication between the feature map and the transpose of the feature map, because the feature map has channel dimension, which is equivalent to the point multiplication of each pixel and every other element. The geometric meaning of the point multiplication of the vector to calculate the similarity of the two vectors. The more similar the two vectors are, the greater the point multiplication is. The enhanced feature is obtained by

multiplying the feature map transpose and the matrix of the feature map and normalizing it with SoftMax. The enhanced feature is then multiplied by the matrix of the transpose of the feature map, which redistributes the correlation information to the original feature map. Finally, the correlation information is added to the initial feature to get the final output, which combines the correlation results of the whole map. In general attention mechanism, an attention map calculates the similarity between all pixels and all pixels, and the spatial complexity is $(H \times W) \times (H \times W)$. In this paper, criss-cross attention is used to calculate the similarity between each pixel and the pixels on the same row or column, that is, the pixels on the cross. By repeating the cycling operation twice, the similarity between pixel and pixel is calculated indirectly. Then, the spatial complexity is reduced to $(H \times W) \times (H + W - 1)$. The criss-cross attention module can effectively improve the global feature without significantly affecting the tracking rate. The details of the criss-cross attention module are shown in Fig. 2.

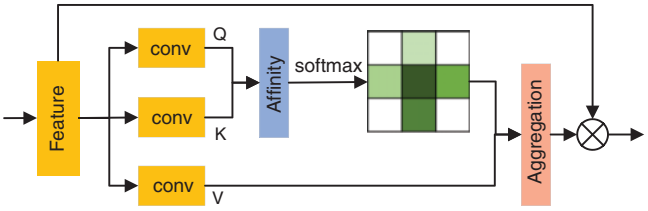

Figure 2: The details of the criss-cross attention-based feature enhanced module

As shown in Fig. 2, the local feature $R^{c\times w\times h}$ is extracted by Semantic network, and two convolutional layers and 1×1 filters are applied to criss-cross attention module H to generate two feature maps Q and$K$, where $\{Q, K\} \in R^{C'\times W\times H}$ is less than C, which is the number of channels of element maps for dimension reduction. After obtaining feature maps Q and K, attention maps $A \in R^{(H+W-1)\times W\times H}$are further generated by attention mechanism operation. At each position u in the spatial dimension of the feature mapping Q, the vector $\hat{Q}_u \in R^{C'}$ can be obtained. At the same time, the Ω_u can be obtained by extracting the eigenvectors from K in the same row or column of the position u, as shown in Eq. (2) below.

$$\Omega_u \in R^{(H+W-1)\times C'} \tag{2}$$

where $\Omega_i, u \in R^{c'}$ is the i-th element of u. Close-relationship operations are defined in Eq. (3) as follows:

$$d_{i,u} = Q_U \Omega_i, u^T \tag{3}$$

where d_i denotes the degree of correlation between feature Q_U and Ω_i. Aggregate operations collect remote contextual information, as follows in Eq. (4):

$$\mathrm{H}'_u = \sum_{i\in|\Phi_u|} A_{i,u}\Phi_{i,u} + H_u \tag{4}$$

where H'_u denotes the eigenvectors in the output feature map $H' \in R^{C'\times W\times H}$ at position u. $A_{i,u}$ is the scalar values of channel i and the positions u of channel A, respectively. The contextual information is added to local feature H to enhance the local features and augment the pixel-wise representation. Therefore, it has a wide contextual view and selectively aggregates contexts according to the spatial attention map.

A criss-cross attention module can capture long-distance contextual information horizontally and vertically, but the connection between pixels and surrounding pixels is still sparse. Feature

enhancement helps obtain dense contextual information. Therefore, we introduce the repeated criss-cross attention module, which can be expanded to R cycles. In the first loop, the criss-cross attention module takes the input feature map H extracted and the output feature map H' from the CNN model as input, where H and H' have the same shape. In the second loop, the criss-cross attention module takes previous input feature mapping H' and output feature mapping H'' as inputs.

As shown in Fig. 3, u and θ are two-position output feature mappings, x and y are the transverse and longitudinal coordinates of u and θ, A_i is the weight of contextual, (x, y) to (x', y') function mapping is expressed as $A_{(i,x,y)} = f(A, x, y, x', y')$. When calculating the global character of the blue pixels in the upper right corner, one-time criss-cross attention can get the correlation of the horizontal and vertical column of pixels (such as blue pixels and green pixels), but it cannot get the correlation of two oblique diagonal pairs of pixels (such as red pixels and blue pixels). Therefore, we use the multiple criss-cross attention modules to transmit and calculate the correlation between red and blue pixels by the information of the two green pixels. The criss-cross attention module of feature enhancement has two loops (R = 2), which is sufficient to obtain remote dependencies from all the pixels to generate new feature mappings with dense and rich contextual information. Compared with the criss-cross attention module, the recurrent criss-cross attention module (R=2) does not bring additional parameters, and can achieve better performance through a smaller calculation increment.

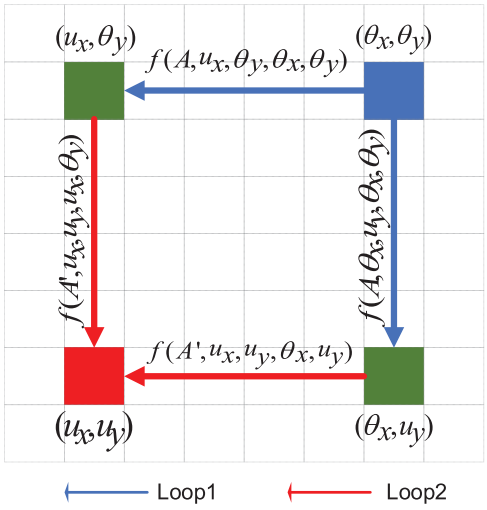

Figure 3: An example of information propagation when the loop number is 2

3.3 Similarity Measurement and Target Area Location

After training, we get the complete SiamCC model and input the video to be tracked into SiamCC. Firstly, we send the template frame Z and the detection frame X into the feature extraction network, respectively. Secondly, we use criss-cross attention on the feature map of the template frame to enhance the global correlation. Then, the similarity matrix is obtained by calculating the similarity of the detected frames with the enhanced feature convolution of the template frames. The similarity function SiamCC is related to a cross-correlation function, which is shown in Eq. (5):

$$f(z, x) = \varphi(x) \times \omega(\varphi(z)) + b1 \tag{5}$$

where $b1$ represents the value of each position in the score map, ϕ represents a feature mapping operation, ω represents feature enhancement operations. SiamCC uses the convolutional layer and

pooling layer in CNN to map the original image to a specific feature space. In SiamCC, features are enhanced by the dilation convolution layer, dimensionality reduction layer, and recurrent criss-cross attention module. The 17×17 similarity matrix is calculated by the cross-correlation function. Finally, it is transformed to 255×255 matrix and 255×255, which is the fitting of the area to be tracked by bicubic interpolation to locate the target region. The principle of bicubic interpolation is that for each pixel x, its pixel value can be obtained by weighting its adjacent left and right two pixels, as shown in Eq. (6):

$$f(x) = \sum_{k=-1}^{2} f(x_k) u(s) \tag{6}$$

where $s = \dfrac{x - x_0}{x_1 - x_0}$ has three times interpolation basis functions for differents values as follows Eq. (7):

$$u(s) = \begin{cases} \frac{3}{2}|s|^3 - \frac{5}{2}|s|^2 + 1 & 0 < |s| < 1 \\ -\frac{1}{2}|s|^3 + \frac{5}{2}|s|^2 - 4|s| + 2 & 1 < |s| < 2 \\ 0 & otherwise \end{cases} \tag{7}$$

3.4 Data Preprocessing and Model Construction in SiamCC

The method uses the image pair for offline training to obtain a training sample pair (z, x) from the annotated video dataset. The image pairs are extracted from two frames of the video, and the maximum distance between the two frames is a fixed value.

The discriminant method is used to train positive and negative samples. The definition of logical loss is shown in Eq. (8):

$$l(y, v) = \log(1 + e^{-yv}) \tag{8}$$

where $y \in (+1, -1)$ denotes the true value, v denotes the actual score of the sample-search image. Eq. (8) denotes the probability of a positive sample $\frac{1}{1+e^{-v}}$ (sigmoid function), and negative sample $1 - \frac{1}{1+e^{-v}}$.

The loss function of Eq. (8) can be easily obtained by the formula of cross-entropy, which is expressed by the average loss of all candidate positions in training:

$$L(y \cdot v) = \frac{1}{|D|} \sum_{u \in D} l(y[u], v[u]) \tag{9}$$

In Eq. (9), D denotes the final score map and u denotes all positions in the score map. The convolutional parameter θ of training is obtained by minimizing the following problems through SGD:

$$\arg \min_{\theta} = E_{(z,x,y)} L(y, f(z, x; \theta)) \tag{10}$$

The determination of positive and negative samples of the network will be output. In the input search image (e.g., 255×255), if the distance between the sample and the target is not more than R, it is a positive sample. Otherwise, it is a negative sample.

$$y[u] = \begin{cases} +1 & if\ k\|u - c\| \le R \\ -1 & otherwise \end{cases} \tag{11}$$

where k is the total step length of the network, c is the center of the target, u is all the positions of the score map, and R is the defined radius.

4 Experiments

4.1 Experimental Dataset

The training dataset is 11668 videos in GOT-10k . The main characters of the videos are moving objects in the real world. The boundary boxes of objects are all hand-marked, totaling more than 1.5 million. The GOT-10k dataset is built based on the WordNet English vocabulary database. It consists of five major categories: animals, artifacts, persons, natural objects, and parts. Five major categories comprise a total of 563 categories. To make the training model have stronger generalization ability, we use GOT-10k as the training set and OTB100 as the test set. The universality and fairness of the algorithm are improved effectively.

4.2 Fine-tuning Retraining in SiamFC

We analyze 9 sequences with poor tracking effect of original SiamFC in OTB100. One of which is target interferer impact tracking. These sequences are divided into two categories. One is the target-distractor influence tracking. Another is the first frame with complex light and shadow. The original SiamFC paper shows that the training data set has a great influence on the tracking accuracy. After the finetuning of parameters on the original SiamFC, we re-trained SiamFC with GOT-10k. As shown in Figs. 4 and 5, the overall accuracy is increased by 3%. The accuracy of the target interferer sequence is greatly improved. However, the poor effect in complex shadow environment is not improved.

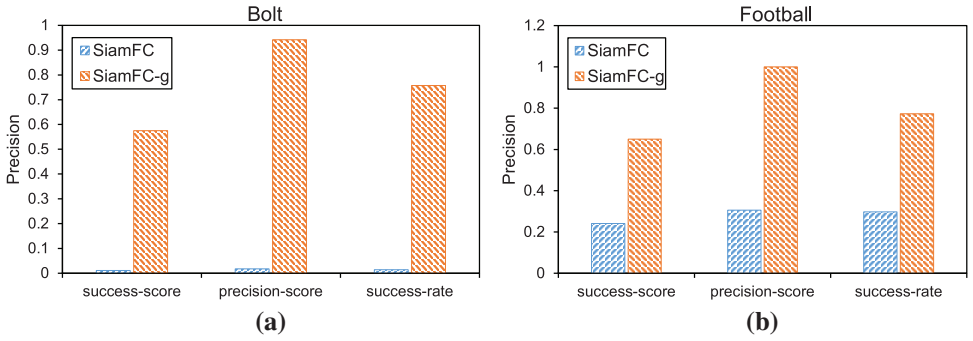

Figure 4: SiamFC-g is a fine-tuned network. The accuracy of target interferer sequences such as Bolt (a) and Football (b) are greatly improved

Figure 5: (Continued)

(b) (d)

Figure 5: The tracking performance of SiamFC and SiamFC-g was demonstrated and compared in similar interferometer sequences

In Fig. 5, (a) is the effect of SiamFC on bolt sequence tracking, (b) is the effect of SiamFC-g on bolt sequence tracking, (c) is the effect of SiamFC on football sequence tracking, and (d) is the effect of SiamFC-g on football sequence tracking. The green box is the SiamFC tracking result, the red line is the truth box, and the blue box is the SiamFC-g tracking result. It can be seen that the accuracy of SiamFC-g is much greater than that of SiamFC.

4.3 The Experiment Details of SiamCC

We also evaluate SiamCC on OTB100. The results of tracking are shown in Figs. 6–10. The tracking effect of SiamCC is greatly improved under the complex shadow environment in the first frame and intersection environment with target interferers.

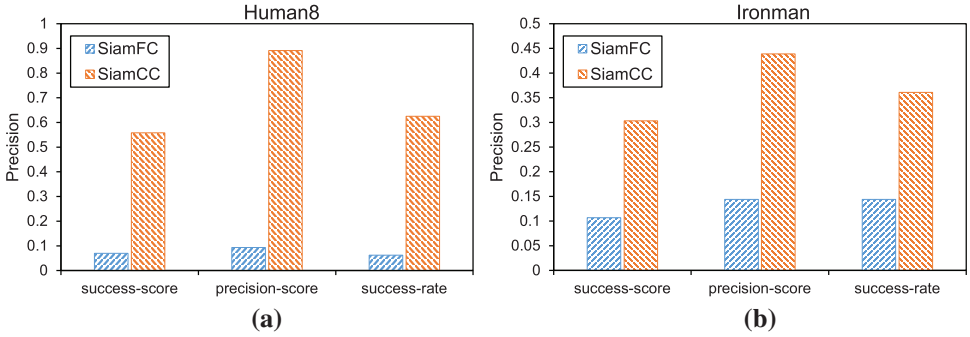

Figure 6: Tracking results of SiamFC and SiamCC in HUMAN8 and IRONMAN sequences under shaded environments

In Fig. 7, (a) is the demo frame of SiamFC for HUMAN8 sequence tracking, (b) is the demo frame of SiamCC for HUMAN8 sequence tracking, (c) is the demo frame of SiamFC for IRONMAN sequence tracking, and (d) is the demo frame of SiamCC for IRONMAN sequence tracking. The tracking results of our method (SiamCC) are compared with the original method (SiamFC) under poor shadow and light conditions. Our algorithm is much more accurate than SiamFC in the face of large changes in illumination and shadow.

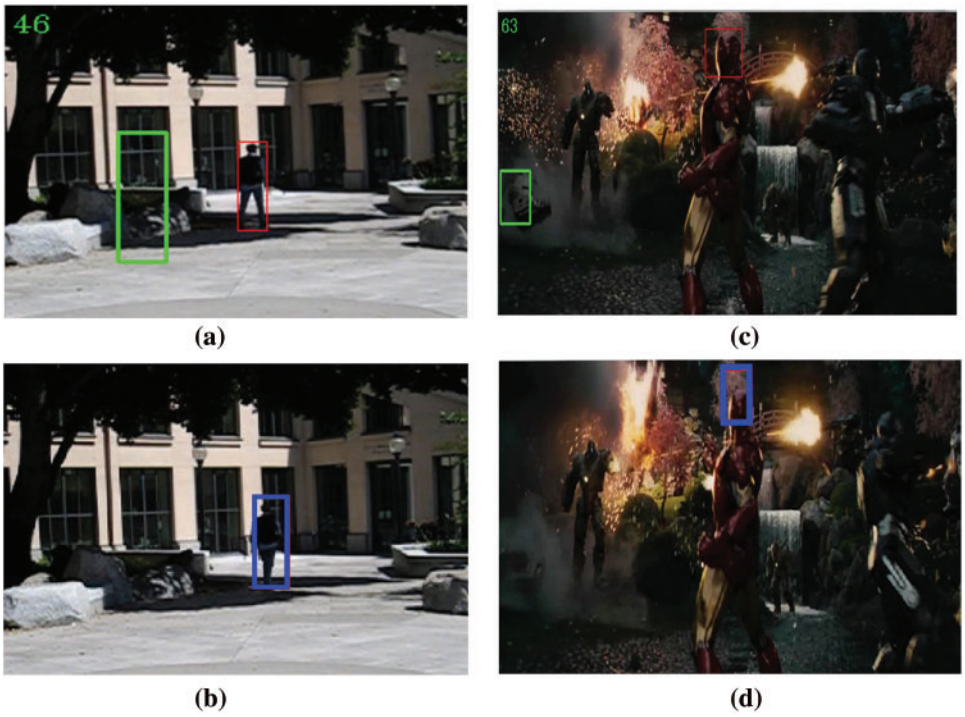

Figure 7: Tracking demo of the SiamFC and SiamCC algorithms in HUMAN8 and IRONMAN sequences under poor shadow and light conditions

In Fig. 8, The red box is the true value, the green box is the tracking box of SiamFC, and the blue box is the tracking box of SiamCC. Since people are not obvious in shadows, SiamFC cannot effectively distinguish shadows from people. SiamCC pays more attention to distinguishing between characters and background environments, so it can better track characters in shadowed environments.

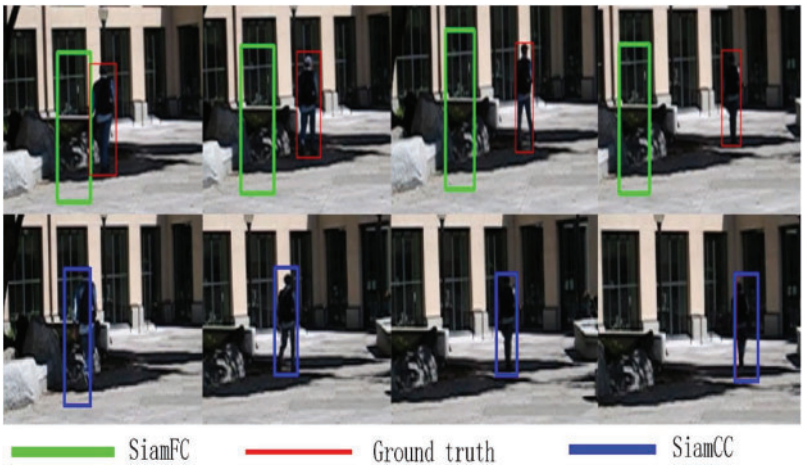

Figure 8: Tracking effects of our trackers SiamCC and SiamFC in shaded environments, respectively. The first row is SiamFC, and the second row is SiamCC

In Fig. 9, The test results on the JUMP and HUMAN3 sequences show that SiamCC is better than SiamFC in terms of tracking accuracy, success rate and score.

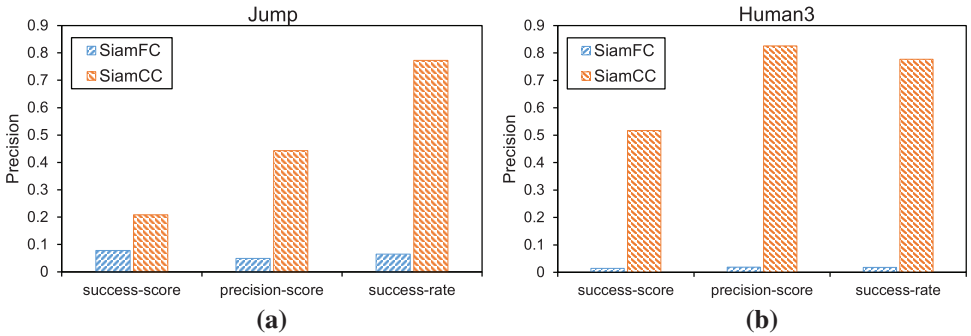

Figure 9: Tracking results of SiamFC and SiamCC in JUMP and HUMAN3 sequences in shaded environments

In Fig. 10, (a) is the demo frame of SiamFC for HUMAN3 sequence tracking, (b) is the demo frame of SiamCC for HUMAN3 sequence tracking, (c) is the demo graph of SiamFC for JUMP sequence tracking, and (d) is the demo frame of SiamCC for JUMP sequence tracking. The tracking results of our method(SiamCC) are compared with the original method(SiamFC) under a similar interference and target intersection environment. From the figure, it can be found that our algorithm is far more accurate than SiamFC in the face of similar interference and target intersection environment.

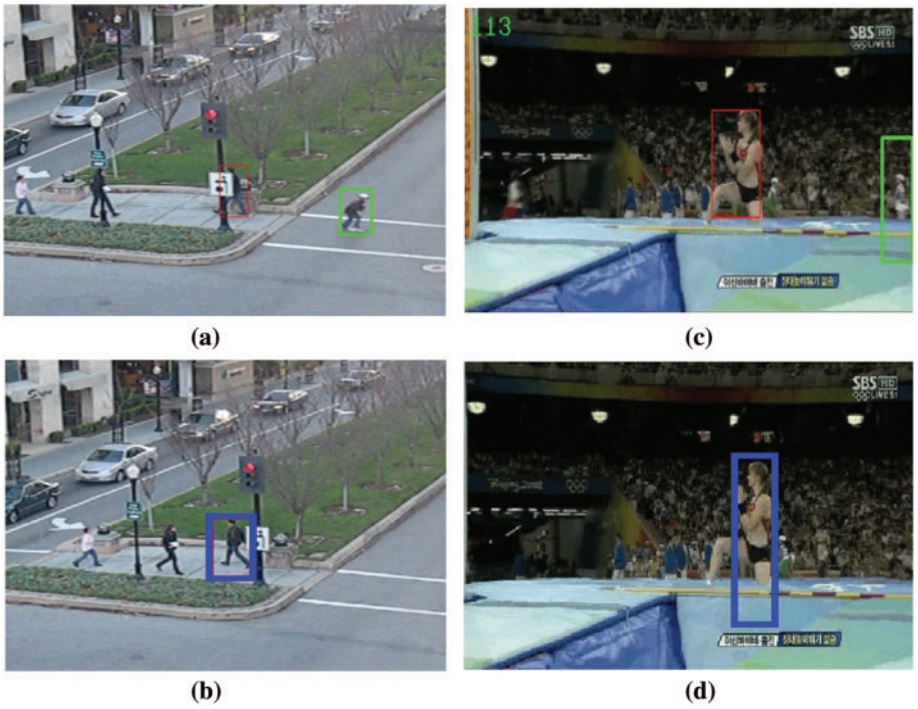

Figure 10: Tracking results of SiamFC and SiamCC in JUMP and HUMAN3 sequences under shaded environments

In SiamCC, we use SiamFC to improve the feature extraction network of AlexNet. The parameters of the first three convolutional layers are fixed, the latter two convolutional layers are fine-tuned, and the dilation convolution is added to the latter. In this method, SGD is used to optimize the loss function. 50 cycles are trained, and the learning speed decreases from 10^{-2} to 10^{-6}.

The training environment is configured as Pytorch 0.4.0 or 0.4.1, 1∗16G GPU, Python 3.6, GCC 4.8.5, and CUDA 8.0. The tracking estimate is done with the GOT-10k Python Toolkit. Since the GPU is only 1050 (2g), the frame rate of the evaluation result is low.

In the process of reproducing CCnet and introducing the criss-cross attention module, we find that the H∗W of feature mapping, the number of repetitions of the criss-cross attention module, and the size of the feature map of the input criss-cross attention module affect each other. Although SiamCC has made some progress in the complex environment, its overall performance fails to surpass the existing algorithms. Therefore, we will further analyze the relationship between them and improve the performance.

5 Conclusion

In this paper, we propose a SiamCC network, which uses a large-scale image pair in GOT-10K dataset for offline training. SiamCC strengthens the object feature attributes by optimizing the global feature correlation in the feature map. Our method can run at 60FPS and achieve advanced performance against real-time challenges in the OTB100 dataset with poor shadow, dim lighting conditions, and a large number of target disturbances.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 62002392, author Y. T, http://www.nsfc.gov.cn/; in part by the Natural Science Foundation of Hunan Province (No.2020JJ4140), author Y. T, http://kjt.hunan.gov.cn/; and in part by the Natural Science Foundation of Hunan Province (No. 2020JJ4141), author X. X, http://kjt.hunan.gov.cn/; in part by the Postgraduate Excellent teaching team Project of Hunan Province under Grant [2019] 370-133, author J. Q, http://xwb.gov.hnedu.cn/; and in part by the Postgraduate Scientific Research Innovation Project of Hunan Province under Grant CX20210878, author Z. W, http://jyt.hunan.gov.cn/; and in part by Scientific Innovation Fund for Post-graduates of Central South University of Forestry and Technology under Grant CX202102056, author Z. W, https://jwc.csuft.edu.cn/.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. Ma, T. Zhou, J. Qin, Q. Zhou and Z. Cai, "Joint-attention feature fusion network and dual-adaptive NMS for object detection," *Knowledge-Based Systems*, vol. 241, no. 2, pp. 108213, 2022.

[2] Z. Wang, J. Qin, X. Xiang and Y. Tan, "A privacy-preserving and traitor tracking content-based image retrieval scheme in cloud computing," *Multimedia Systems*, vol. 27, no. 3, pp. 403–415, 2021.

[3] C. Wang, Y. Liu, Y. Tong and J. Wang, "GAN-GLS: Generative lyric steganography based on generative adversarial networks," *Computers, Materials & Continua*, vol. 69, no. 1, pp. 1375–1390, 2021.

[4] L. Xiang, J. Qin, X. Xiang, Y. Tan and N. N. Xiong, "A robust text coverless information hiding based on multi-index method," *Intelligent Automation & Soft Computing*, vol. 29, no. 3, pp. 899–914, 2021.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause and S. Satheesh, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2014.

[6] L. Xiang, X. Shen, J. Qin and W. Hao, "Discrete multi-graph hashing for large-scale visual search," *Neural Processing Letters*, vol. 49, no. 3, pp. 1055–1069, 2019.
[7] J. Zhang, X. Jin, J. Sun, J. Wang and K. Li, "Dual model learning combined with multiple feature selection for accurate visual tracking," *IEEE Access*, vol. 7, no. 1, pp. 43956–43969, 2019.
[8] Q. Liu, X. Xiang, J. Qin, Y. Tan, J. Tan *et al.,* "Coverless steganography based on image retrieval of DenseNet features and DWT sequence mapping," *Knowledge-Based Systems*, vol. 192, no. 1, pp. 105375–105389, 2020.
[9] Q. Zhou, J. Qin, X. Xiang, Y. Tan and Y. Ren, "MOLS-Net: Multi-organ and lesion segmentation network based on sequence feature pyramid and attention mechanism for aortic dissection diagnosis," *Knowledge-Based Systems*, vol. 239, no. 17, pp. 107853, 2021.
[10] Y. Luo, J. Qin, X. Xiang and Y. Tan, "Coverless image steganography based on multi-object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2779–2791, 2021.
[11] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 4293–4302, 2016.
[12] R. Tao, E. Gavves and A. W. Smeulders, "Siamese instance search for tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 1420–1429, 2016.
[13] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 850–865, 2016.
[14] Q. Wang, Z. Teng, J. Xing, J. Gao and W. Hu, "Learning attentions: Residual attentional Siamese network for high performance online visual tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 4854–4863, 2018.
[15] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 5000–5008, 2017.
[16] Q. Guo, W. Feng, C. Zhou, R. Huang and L. Wan, "Learning dynamic Siamese network for visual object tracking," in *Proc. Int. Conf. on Computer Vision*, Honolulu, Hawaii, USA, pp. 1781–1789, 2017.
[17] B. Li, J. Yan, W. Wu, Z. Zhu and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 8971–8980, 2018.
[18] Z. Zhu, Q. Wang, B. Li, W. Wu and J. Yan, "Distractor-aware Siamese networks for visual object tracking," in *Proc. European Conf. on Computer Vision*, Salt Lake City, Utah, USA, pp. 103–119, 2018.
[19] H. Fan and H. B. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 7944–7953, 2019.
[20] Q. Wang, L. Zhang, L. Bertinetto, W. M. Hu and P. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 1328–1338, 2019.
[21] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing *et al.,* "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 4277–4286, 2019.
[22] J. Wang, J. Qin, X. Xiang, Y. Tan and N. Pan, "CAPTCHA recognition based on deep convolutional neural network," *Mathematical Biosciences and Engineering*, vol. 16, no. 5, pp. 5851–5861, 2019.
[23] S. Zhao, M. Hu, Z. Cai and F. Liu, "Dynamic modeling cross-modal interactions in two-phase prediction for entity-relation extraction," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021. https://doi.org/10.1109/TNNLS.2021.3104971.
[24] L. Huang, X. Zhao and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the Wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1562–1577, 2021.

[25] Y. Li, Y. Q. Guo, Y. Y. Kao and R. He, "Image piece learning for weakly supervised semantic segmentation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 648–659, 2017.
[26] S. Zhao, M. Hu, Z. Cai, Z. Zhang, T. Zhou *et al.,* "Enhancing chinese character representation with lattice-aligned attention," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021. https://doi.org/10.1109/TNNLS.2021.3114378.
[27] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7132–7141, 2018.
[28] L. Chen, Y. Yang, J. Wang, W. Xu and A. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 3640–3649, 2016.
[29] W. Pan, J. Qin, X. Xiang, Y. Wu, Y. Tan *et al.,* "A smart mobile diagnosis system for citrus diseases based on densely connected convolutional networks," *IEEE Access*, vol. 7, pp. 87534– 87542, 2019.
[30] X. Wang, R. Girshick, A. Gupta and K. He, "Non-local neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7794–7803, 2018.
[31] J. Fu, J. Liu, Y. Li, Y. Bao and H. Tian, "Dual attention network for scene segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 3141–3149, 2018.
[32] H. Zhao, Y. Zhang, S. Liu, J. Shi and C. C. Loy, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. the European Conf. on Computer Vision*, Munich, Germany, pp. 270–286, 2018.
[33] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei *et al.,* "CCNet: Criss-cross attention for semantic segmentation," in *Proc. the IEEE Conf. Int. Conf. on Computer Vision*, Seoul, Korea, pp. 603–612, 2019.
[34] J. Zhang, C. Lu, X. Li, H. J. Kim and J. Wang, "A full convolutional network based on DenseNet for remote sensing scene classification," *Mathematical Biosciences and Engineering*, vol. 16, no. 5, pp. 3345–3367, 2019.
[35] Y. Wu, J. Lim and M. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.
[36] T. Zheng, Y. Luo, T. Zhou and Z. Cai, "Towards differential access control and privacy-preserving for secure media data sharing in the cloud," *Computers & Security*, vol. 113, no. 1, pp. 102553, 2022.
[37] Z. Zhou, J. Qin, X. Xiang, Y. Tan, Q. Liu *et al.,* "News text topic clustering optimized method based on TF-IDF algorithm on spark," *Computers, Materials & Continua*, vol. 62, no. 1, pp. 217–231, 2020.
[38] J. Zhang, C. Lu, J. Wang, L. Wang and X. Yue, "Concrete cracks detection based on FCN with dilated convolution," *Applied Sciences*, vol. 9, no. 13, pp. 2686, 2019.