Anchor-free Object Tracking Algorithm Combining RFB and Dual Attention

Yucheng Wang*
School of Computer Science, Wuhan University, Wuhan, Hubei, China
**Corresponding Author*

Abstract

In view that the Siamese trackers could not capture the long-range dependence and vulnerable to problems similar to objective factors such as interference of background, this paper proposes a lightweight dual attention module, it can be increased under the condition of less amount of calculation to efficiently capture the attention of space dimension and the channel dimension. Meanwhile, we introduced the RFB module to strengthen the feature. Specifically, our method is mainly divided into two steps in the feature fusion stage: (i) Use the RFB module to strengthen the single-scale template features and search area features. (ii) Use the dual attention module to make the network focus on the distinguishing and robust features of the target adaptively to eliminate similar targets' interference. In the dual attention module, this study uses cascaded two Criss-Cross attention modules to model spatial dimension attention. Extensive experiments on OTB2015 and LaSOT datasets show that this object tracking algorithm embedded with the RFB module and the lightweight dual attention module not only achieves good performance compared to the most advanced tracking algorithms and still runs at a real-time speed of 45FPS.

Keywords*: Object tracking, Siamese network, dual attention, RFB*

I. Introduction

Object tracking is an essential task in the field of computer vision. The tracker needs to accurately predict the target position and size change in the subsequent video frames according to the target object given in the first frame of the video sequence. Although much progress has been made in the study of object tracking [1], it is still a challenging task because the tracking target is often disturbed by external factors such as size change, illumination change, and occlusion [2].

In recent years, most single object tracking algorithms use convolution neural network to extract the target's deep features. The Siamese network based object tracking algorithm uses the deep convolution neural network with shared parameters to extract the features of the target template and the search area respectively. Through cross-correlation operation [3], the response map can be obtained, representing the similarity between the target and different positions in the search area. In order to determine the specific position and shape of the target in the search area according to the response map, the current tracking algorithms are divided into full-convolutional Siamese tracking algorithms, region proposal or anchor box based Siamese neural network tracking algorithm, and the object tracking algorithm without anchor box. The full-convolutional Siamese tracking algorithms, such as SiamFC [3] and SiamDW [4], directly map the target position through the peak value of the response map. The Siamese tracking algorithm based on region proposal or anchor box, such as SiamRPN [5], needs to define the number of anchor boxes in advance and determine the location of the tracked target through classification and regression of anchor boxes. Based on the object tracking algorithm without anchor box, such as Ocean [6], such algorithm does not need the pre-set anchor box, but directly predicts the position of the tracked target. It can be seen that the above algorithms all carry out target location and shape regression prediction after feature extraction and fusion of template area and search area. However, when there is confusion or background clutter factors, such as feature extraction, these factors will hinder the tracker for the distinguishable characteristics of the tracked target has to learn.

In order to solve the Siamese trackers for the position of an important degree of different characteristics of the picture

and channel gives the same weight, but also unable to capture information over long distances, attention mechanism [7, 8] can learn directly to the characteristics of the weight of different locations and different channels of information in the chart, to get the goal of more robust features, and ignore information has nothing to do with the target. Inspired by non-local blocks, Woo et al. [9] proposed CPBAM to generate attention maps based on space and channel respectively, and stack them sequentially. Fu et al. [10] designed Dual attention Network to capture spatial and channel attention respectively. Cao et al. [11] proposed that a global context module was designed based on the non-local module and the extrusion excitation module, which significantly reduced the computational complexity and the number of parameters.

In recent years, attention mechanism has been applied to the field of object tracking. Zhang and Wang [12] processed the feature map extracted by convolution neural network through full convolution dual correlation filter and non-local module to get the fused response map. In order to map the similarity between the target and the background, Liu et al. [13] introduced the non-local function and fused the non-local information to form the target feature. DensSiam [14] introduced non-local modules to make the network learn more robust features of the tracked target. All the above methods use non-local attention mechanism to capture remote information, so they are more robust to the changes in the target's appearance. However, due to the introduction of non-local modules in these methods, the model parameters are significantly increased, which seriously affects the speed of the object tracking algorithm. At the same time, these methods only model attention in spatial dimension, ignoring attention relationship in channel dimension. In this paper, we propose an efficient Dual attention module for object tracking. This module integrates Dual attention Network and Criss-Cross attention Module to extract spatial and channel attention maps and reduce the number of parameters. We embed the module into the feature fusion stage of the Ocean to solve the problem of unable to capture the characteristics of long-distance information. After this module, the fused feature map can highlight the target location when there are interference factors in the background or occlusion of the target itself and make the tracking algorithm robust to the shape change of the target. At the same time, we introduced the RFB [15] module to strengthen the single scale feature before the depth-wise cross-correlation operation in the feature fusion stage. Firstly, multi branch convolution kernel with different sizes is used to extract the features of receptive fields with different sizes. Then dilation convolution with different dilation rates is used for each branch to highlight more critical information in each receptive field, so the module can effectively enhance the features. Meanwhile, the calculation cost of this module is small, and the real-time performance of the original tracking algorithm will not be affected.

Our main contributions can be summarized as follows.

We propose an efficient Dual attention module that can extract spatial and channel attention and reduce the number of parameters required by the module. The fused feature map after this module can highlight the target location when there are interference factors in the background or occlusion of the target itself.

We introduced the RFB [15] module to strengthen the single scale feature instead of the dilated convolution layer. The RFB module can effectively enhance the features before the depth-wise cross-correlation operation in the feature fusion stage.

We use Dual attention module and RFB module to enhance the robustness of the object tracking algorithm to the shape change of the target. At the same time, due to the small amount of computation of these two modules, we make the speed of the tracking algorithm still reach 43fps, which fully meets the real-time requirements.

This research conducted comprehensive experiments on the improved algorithm proposed in this paper on two challenging data sets, OTB2015 [16] and LaSOT [17]. Compared with the baseline algorithm Ocean, this method achieves great improvement on both data sets. In addition, this algorithm also achieves the leading performance by comparing with the current mainstream tracking algorithms.

II. Related Work

2.1 Attention mechanism

The mechanism of attention improves the accuracy of viewing specific areas by imitating the internal processes of biological observing objects and by combining global information to assign different weights to different locations. However, since the use of stacked convolutional layers to capture remote information requires a high computational cost and is difficult to be optimized, and information between remote pixels is difficult to be transmitted under such a structure, the current attention mechanisms that can capture remote information are mainly divided into the following two types:

Self-attentional mechanism. In recent years, self-attention mechanism has been applied in many fields, such as machine translation [18] and computer vision tasks [19-21], because it reduces the dependence on external information and is better at capturing the internal correlation of features. Literature [18] firstly applied the self-attention mechanism to long-distance information modeling in machine translation. The non-local module [7] generates an attention map for each query location of the feature map in the way of self-attention. CCNet [22] accelerates the calculation of non-local modules by stacking two Criss-Cross attention modules and successfully applies it to semantic segmentation.

Channel attention mechanism. In order to learn the weight information corresponding to different channels, SENeT [7] and PSANeT [23] learn the weight relationship between different channels in the feature map by rescaling different channels. CBAM [9] perform rescaling to learn the weight relationship of different spatial positions and channels. The Global Context (GC) Block integrates the Senet and the non-local module. It simplifies the autocorrelation operation in the non-local module by learning a shared attention diagram for all the query positions in the feature map, and adopts a method similar to the Senet to obtain the attention between channels.

2.2 Receptive field

Since convolution kernels of different sizes can capture the information of receptive fields of different sizes, the identification and robustness of features can be enhanced by combining convolution kernels of different sizes. Current studies on receptive fields in the field of CNN include Inception Block [24], ASP [25], Deformable CNN [26], and RFB, respectively.

Inception Block uses multiple convolutional kernel branches of different sizes to capture multi-scale information. However, these convolutional cores are all based on the same center for sampling. ASPP (Atrous Spatial Pyramid Pooling) uses expansion convolution to change the sampling distance from the center, but only a uniform size convolution kernel is used, making it unable to adapt to the change of target size. DCN (Deformable CNN) is adapted according to the spatial distribution (scale and shape) of the target receptive field, but the center of the receptive field is not taken into account, the contribution of all pixels in the receptive field is the same, and some critical information is not strengthened. RFB firstly captures multi-scale information through convolution kernels of different sizes and then uses expansion convolution with different dilated rates to strengthen important information in each receptive field. Therefore, RFB can effectively strengthen feature information to improve its recognizability and robustness.

III. Method

3.1 Overall framework of algorithm

The algorithm in this paper mainly includes three parts: feature extraction, feature fusion, and anchor-free classified-

regression network. Its overall structure is shown in Fig 1.

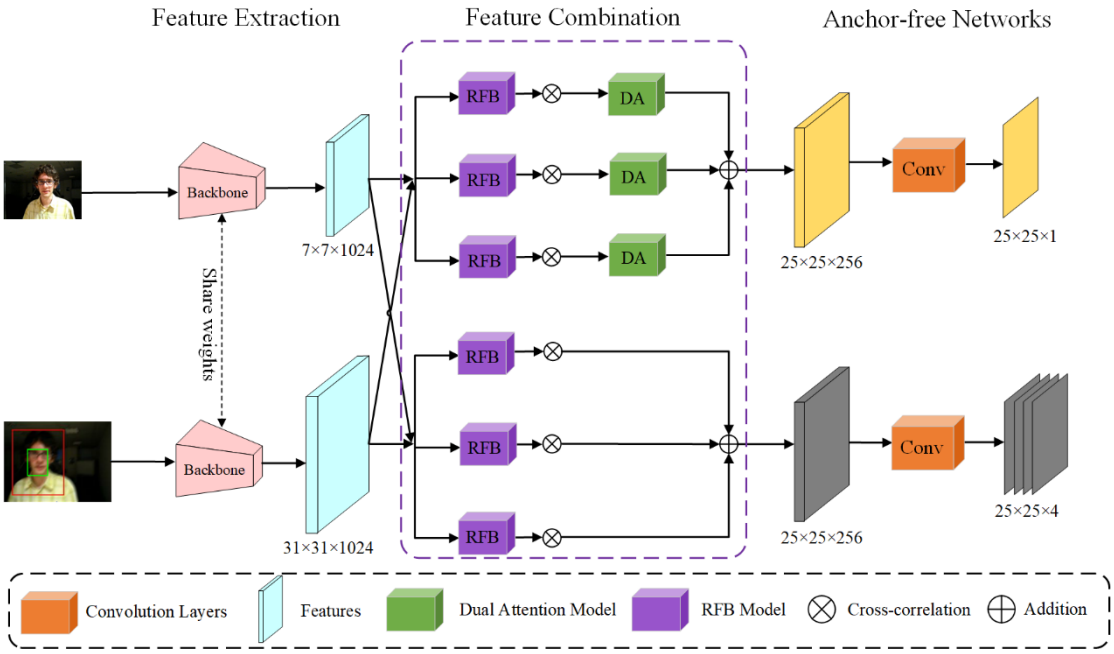

Fig 1: Overall structure of the algorithm for this study

In the feature extraction part, the image pairs formed by the template area and the search area are taken as input. The template area is the image block centering on the tracking target in the first frame of the video sequence, while the search area is the area where the tracking target may exist in the subsequent frames. Then, the input images were extracted from the backbone with shared parameters, which only retained the first four stages of RESNET-50 [27]. In the fourth stage, the convolution step size of the lower sampling unit was reduced from 2 to 1, and the step size of 3×3 convolution was increased to 2. Finally, this study can get the feature map of the template area and the search area, respectively.

In the feature fusion part, as shown in Fig 1, we first pass the two feature maps obtained from the feature extraction part through three parallel RFB modules in the classification branch and the regression branch, and then conduct depth-wise cross-correlation [18] between the three groups of RFB template region feature maps and the search region feature maps in each branch, Three response graphs can be obtained in the classification branch and the regression branch respectively. Since the purpose of the classification branch is to determine whether each position in the cross-correlation diagram belongs to the tracked target, this study input the three response maps obtained from the classification branch into three parallel dual attention modules to enhance the intensity of the tracked target in the response map and eliminate the interference of similar targets in the background. Finally, the three response maps enhanced by double attention are added element by element to get the response map used for the classification branch. In the regression branch, this study directly carries out the element-by-element phase of the three response maps after the depth-wise cross-correlation to get the response map that is finally used in the regression branch. Specifically, the classification branch and regression branch of feature fusion can be respectively expressed as Equations 1 and 2:

$$S_{cls} = \sum_{i=1}^{3} \sigma_i\left(G_i(\psi_e) \times G_i(\psi_s)\right) \quad (1)$$

$$S_{reg} = \sum_{i=1}^{3}\left(G_i(\psi_e) \times G_i(\psi_s)\right) \quad (2)$$

The S_{cls} and S_{reg} represent the response maps of classification branch and regression branch respectively; ψ_e and ψ_s represent the feature map of template area and search area respectively; $G_i(\cdot)$ Represents the i th RFB module in

each branch;$\sigma_i(\cdot)$ represents the No. i dual attention module in the classification branch.

In the part of the anchor-free classified-regression network, the classification probability map is obtained by passing through 4 standard convolution layers and a convolution layer with the number of convolution cores of 1, which represents the classification confidence of each position in the original classification response map. For the regression response map obtained, a regression coordinate map is obtained by passing through 4 standard convolution layers and then a convolution layer with the number of convolution cores of 4, representing the boundary box corresponding to each position in the regression response map.

3.2 RFB module

The RFB module introduced in this paper is shown in Fig 2. It can be seen that the RFB module is mainly divided into two parts, namely the multi-branch convolution layer with different sizes of convolution cores and the subsequent dilated convolution layer. The multi-branch convolutional layer uses the multi-branch convolution layer with the convolution kernel size of 1×1, 3×3, 5×5 to process the input feature map. The second part is the dilated convolution layer with the corresponding expansion step after each branch convolution layer. The RFB module can capture more context information while reducing the number of new parameters as much as possible.

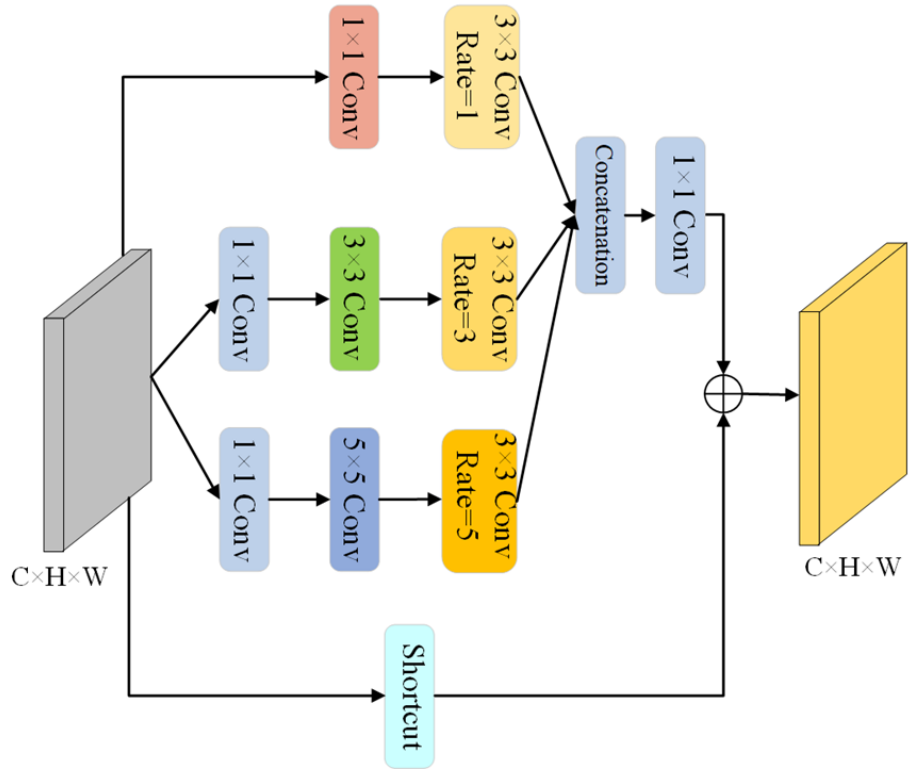

Fig 2: The RFB Module

3.3 Dual attention module

The overall structure of the dual attention module proposed in this paper is shown in Fig 3. It can be seen that, after the feature graph of the input module is strengthened by the position attention module and the channel attention module, respectively, it is fused by adding elements one by one to get the final enhanced feature map. Next, this study will look at the location and channel attention modules in detail.

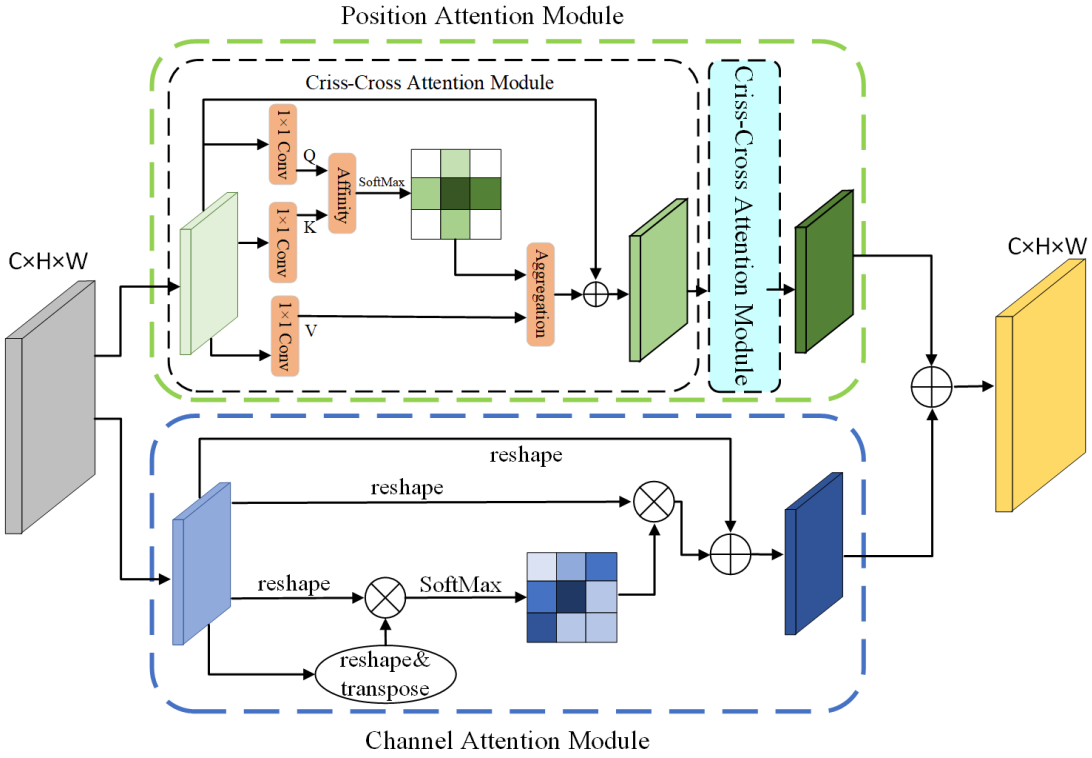

Fig 3: Dual attention module

Spatial attention module: Although the non-local module can capture the information relationship at a distance in space by calculating the similarity of any two positions in the feature map, it consumes a lot of GPU memory and has high computational complexity. Therefore, this study uses the Criss-Cross attention module, which consumes less GPU memory and has lower computational complexity to replace the non-local module to capture spatial attention. For each query position in the input feature map with a dimension of C×H×W, the Criss-Cross attention module generates a weight mask of H+W-1, which represents the correlation degree between each query position and other positions on the crosshairs centering on this position. By cascading the two Criss-Cross attention modules, the remote correlation of all pixels can be captured at each position in the final output feature map.

Channel attention module: Because the channels of each high-level feature are correlated with each other, the feature of each channel can be enhanced by using the mutual information association between channels. This study constructs a channel attention module to capture the information relationship between channels explicitly. Different from the spatial attention module, this study directly calculates the channel attention map $X \in R^{C \times C}$ from the input feature map A whose dimension is C×H×W. To be specific, let us first transform A to $R^{C \times N}$, and then matrix multiplication was performed for A and the transposed A. Finally, Softmax was used to obtain the channel attention diagram $X \in R^{C \times C}$, the process is shown in Equation 3.

$$x_{mn} = \frac{exp(A_m \cdot A_n)}{\sum_{m=1}^{C} exp(A_m \cdot A_n)} \tag{3}$$

Where, x_{ji} is the correlation degree of the m channel to the n channel. Finally, we use Equation 4 to obtain the final output feature map E∈C×H×W.

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j \tag{4}$$

Where β learns the weight from random initialization. It can be seen from formula 4 that the final feature map is obtained by the weighted addition of the features of all channels and the original features. The channel attention module can model the long-term semantic information between feature maps, which is helpful to improve the ability of feature discrimination.

IV. The Experiment

4.1 Algorithm implementation details

The algorithm in this paper is implemented using PyTorch, and the running environment is ubuntu 18. 04, the CPU is Intel Core i9, and the GPU is Nvidia 2080TI.

Firstly, the algorithm in this paper initializes the parameters of the whole network with the parameters of the trained Ocean algorithm. During the training process, the parameters of the backbone network were frozen, while the feature fusion of the classification branch and the regression branch, and the anchor-free network were fine-tuned. We used YouTube-BB [28], GOT-10k [29] and COCO [30] data sets to train the algorithm in this paper. The template image size in the algorithm input image pair is 127×127, and the size of the image in the search area is 255×255. In this paper, the momentum parameter is 0. 9 and the weight attenuation is 10-3. Batchsize is set as 32, and the number of training cycles is 50, in which 4×105 is used for each cycle.

4.2 Experimental data set and evaluation criteria

In order to evaluate the performance of the algorithm in this paper, this study tested the algorithm in this paper on two public datasets, OTC2015 and LaSOT. Among them, the OTC2015 short-time tracking dataset contains 100 video sequences in total and is labeled with 11 attributes, including MB (motion blur), LR (low resolution), BC (background clutter), etc. LaSOT is divided into training set and test set, in which the test set contains 280 test sequences, each test sequence has an average frame number of 2500 frames, and 14 attributes such as CM (camera motion), OV (target out of view), BC (background clutter) are also marked. Compared with OTB2015 data set, LaSOT data set is larger and more challenging. It not only has a longer test sequence, but also has more complex tracking scenes with more interference factors. Therefore, it is easier to have similar targets in the background and targets are occluded. These two kinds of data sets can well test the accuracy and robustness of the proposed algorithm in various complicated cases of short-time tracking and long-time tracking.

Both OTB2015 and LaSOT adopt two evaluation indexes: precision chart and success rate chart. The precision chart represents the percentage of the total frames in which the distance between the center of the target box predicted by the tracking algorithm and the manually marked target box is within 20 pixels. The success chart shows the percentage of the total frames with the predicted target box and the manually labeled target box overlapping rate greater than 0. 5.

4.3 Experimental results of OTB2015 dataset

4.3.1 Overall performance evaluation of the algorithm

This research compares the proposed algorithm with seven mainstream tracking algorithms, namely MDNeT [31], GradNet[32], DiMP [33], ATOM [34], CFNeT [35], SiamFC [3] and the baseline algorithm Ocean. Fig 4 shows the success rate and precision rate of the above 8 algorithms on the OTC2015 dataset. It can be seen from the figure that compared with the baseline algorithm Ocean, the success rate of the proposed algorithm has increased by 0. 9%, reaching 67. 8%, while the precision rate has increased by 0. 88%, reaching 91%.

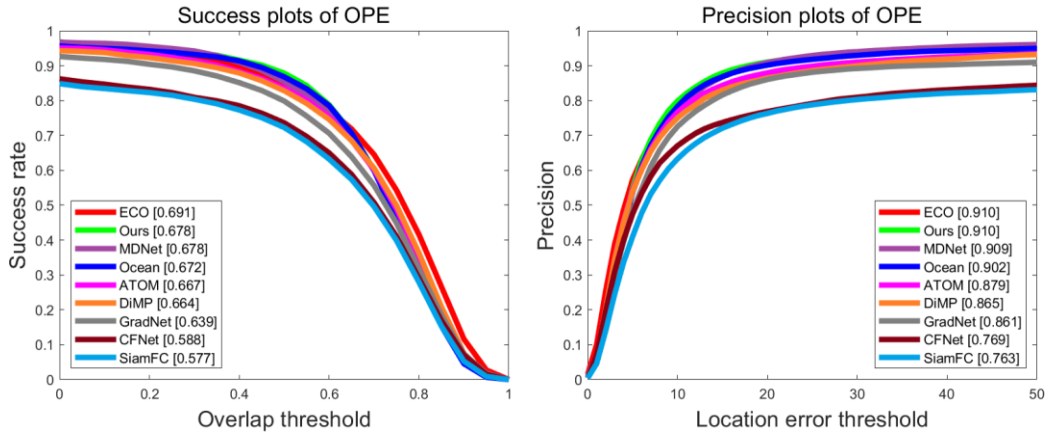

Fig 4: Success plot and Precision plot of OTB2015

4.3.2 Performance evaluation of the algorithm under different attributes

In order to further analysis in this paper, the performance of algorithm under different noise, this study respectively in this paper, the algorithm with the above seven kinds of algorithms in OTB2015 11 kinds of interference factors of the presence of a separate test, because this algorithm was introduced to the baseline algorithm in the Ocean RFB module with dual attention module, therefore, likely to cause a similar target jamming, target deformation and keep out the attribute is that this study need to focus on. According to the attribute of each sequence annotation in OTB2015, this study chose the background clutter (BC), the target is wholly or partially occluded (OCC), these three attributes to target deformation (DEF) analysis of the algorithm in this paper, can be seen from the Fig 5, the algorithm in this paper with the above three test results on the properties of video sequences in the Ocean success rate relative to a baseline algorithm respectively increased by 2.5%, 1.2%, 1.6%, and the precision improved respectively 2.9%, 1.3%, 2.9%.

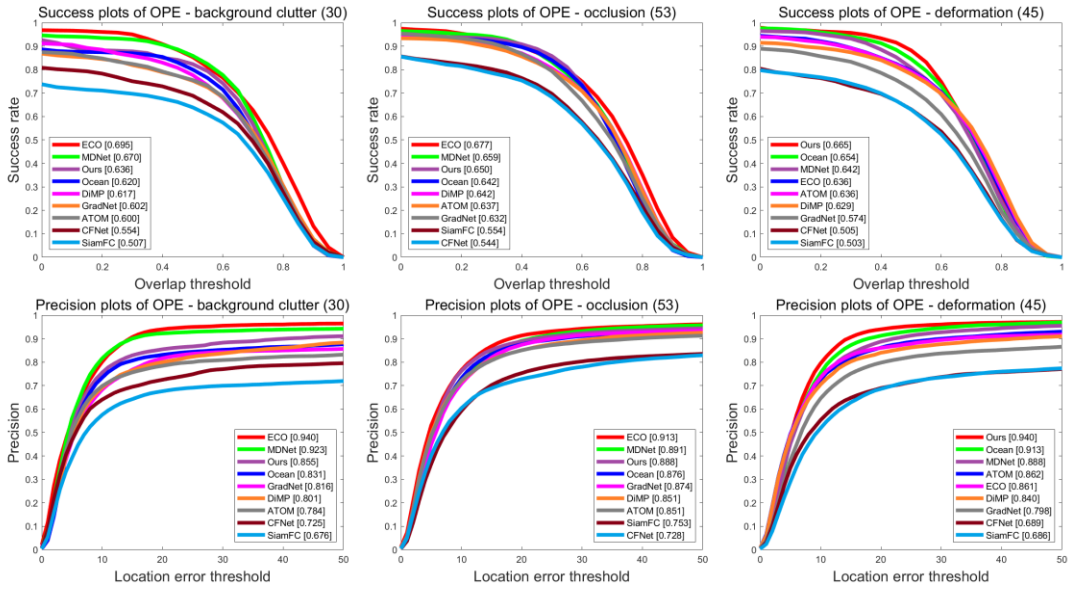

Fig 5: Success plot and Precision plot under three interference conditions in OTB2015

4.3.3 Algorithm qualitative evaluation

In this section, this study selects three video sequences to further analyze the performance of the algorithm proposed. These video sequences contain properties such as interference of similar targets and occlusion of targets. Screenshots of their tracking videos are shown in Fig 6, and the videos from top to bottom are Jogging, Box and Liquor respectively. Due to Ocean algorithm features fusion part only uses a single scale and unable to capture information

over a long distance, so when the background of similar objects to be tracked target will thus resulting in multiple peak response figure, make easy to lose the object tracking algorithm, at the same time, if by tracking target by partial or complete keep out, can lead to Ocean algorithm is unable to accurately extract the distinguishable characteristics of the target, resulting in lost goals. As shown in Fig 6(a), in the Jogging video sequence, the tracked target is completely occluded at frame 78. Meanwhile, since there are characters with similar tracked targets in the background, the Ocean algorithm loses the target due to interference from similar targets after the target is occluded. As shown in Fig 6(b), in the Box video sequence, due to the presence of a large number of square targets similar to the tracked target in the background, the Ocean algorithm lost the target in the 342 frame and was unable to recapture the target in subsequent frames. However, there are many similar glass bottles in Fig 6(c), and the Ocean algorithm misidentifies similar targets in the background as tracked targets after the target is occluded several times. However, due to the addition of RFB module and dual attention module, the algorithm in this paper can not only learn more robust features of the tracked target to resist the interference caused by partial or complete occlusion of the target, but also capture distant information to accurately locate the tracked target when there are similar targets in the background. Therefore, the algorithm in this paper can accurately capture the tracked target in the above video sequence when the target is occluded or there are similar targets in the background.

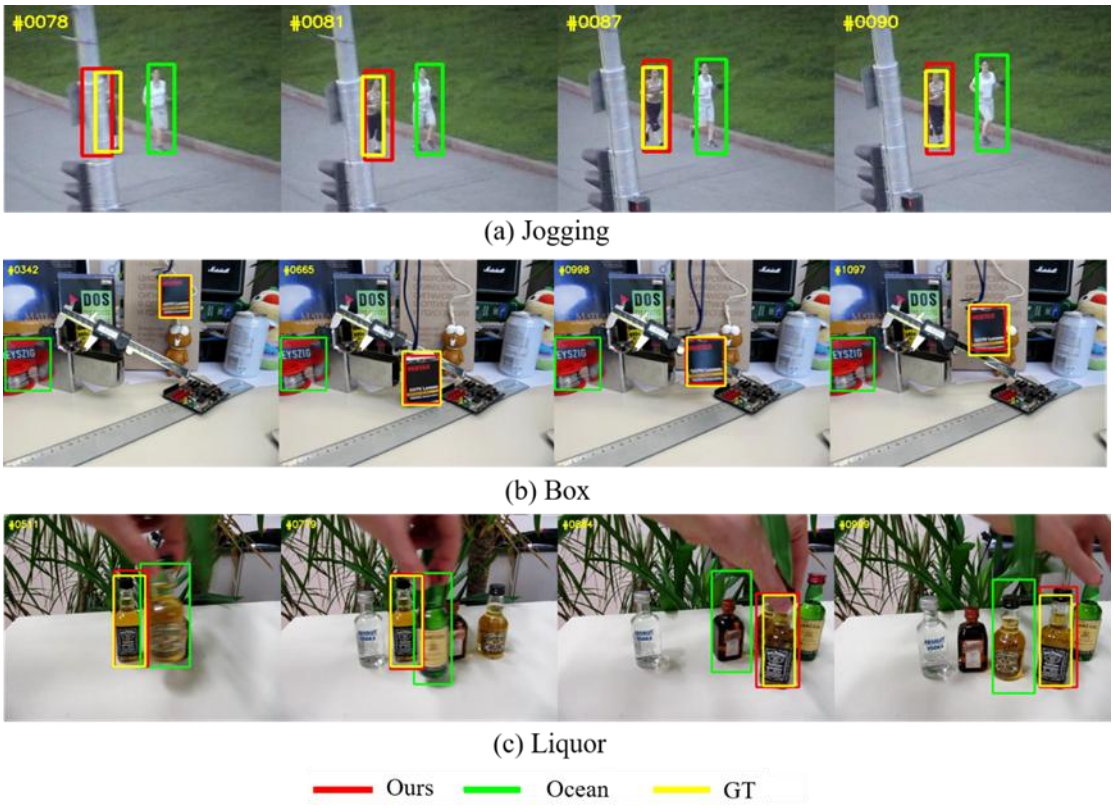

Fig 6: Screenshot of partial tracking video of OTB2015 dataset

4.3.4 Algorithms track speed estimates

In order to evaluate the tracking speed of the algorithm in this paper, this study compared the tracking frames per second (FPS) of the algorithm in this paper and the above six algorithms on the OTC2015 dataset, and the results are shown in Table 1. The test environment is Nvidia 2080TI GPU. It can be seen from the table that although the RFB module and dual attention module are added to the Ocean, the speed of the algorithm in this paper is not greatly reduced compared with the Ocean because the computational cost of the dual attention module and the RFB module are relatively small. The algorithm proposed in this paper can still achieve real-time tracking speed.

Table 1 The Speed of different algorithms on the OTB2015 dataset

Algorithm	Speed (FPS)
GradNet [32]	80
SiamFC [3]	58
Ocean [6]	52
Ours	45
DiMP [33]	43
CFNet [35]	43
ATOM [34]	32
MDNet[31]	5

4.4 Experimental results of LaSOT dataset

4.4.1 Overall performance evaluation of the algorithm

On LaSOT data set, this study compared the algorithm in this paper with other six tracking algorithms, which are ATOM [34], MDNET [31], SiamRPN ++ [18], VITAL [36], ECO [37] and SiamFC [3]. Fig 7 shows the success rate and precision of all the algorithms. It can be seen from the figure that the success rate of the algorithm in this paper reaches 53. 9%, which is 6. 7% higher than that of the baseline algorithm Ocean, while the precision rate reaches 51. 4%, which is 6% higher than that of the baseline algorithm Ocean.

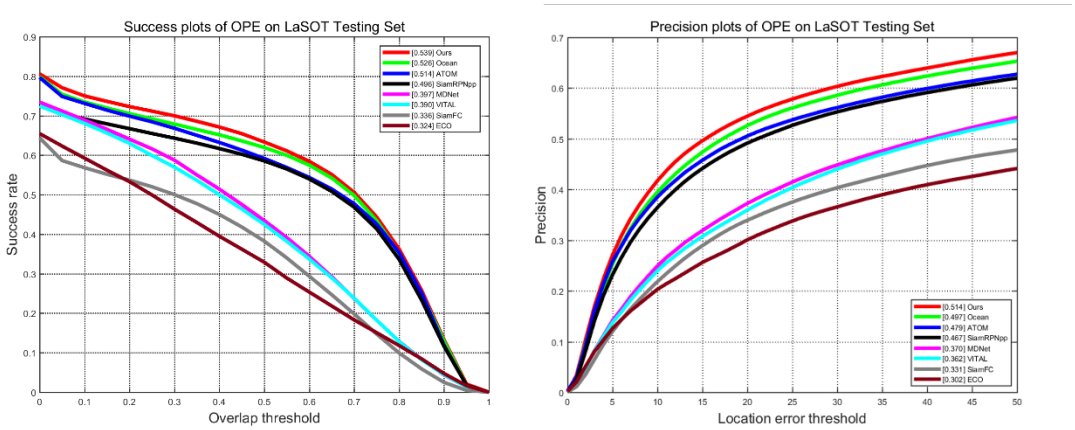

Fig 7: Success plot and Precision plot of LaSOT

4.4.2 Performance evaluation of the algorithm under different attributes

In this section, this study also selects four attributes in LaSOT that may cause similar target interference and target deformation to evaluate the performance of the algorithm in this paper. These four attributes are target partially occluded (POC), target deformation (DEF), background clutter (BC) and target completely occluded (FOC). As can be seen from Fig 8, the success rate of the algorithm in this paper is respectively improved by 3. 7%, 1. 2%, 3. 8% and 1. 5%, and the precision is respectively improved by 5. 3%, 2. 2%, 5. 6% and 2. 2%, compared with the baseline algorithm Ocean under the above 4 attributes.

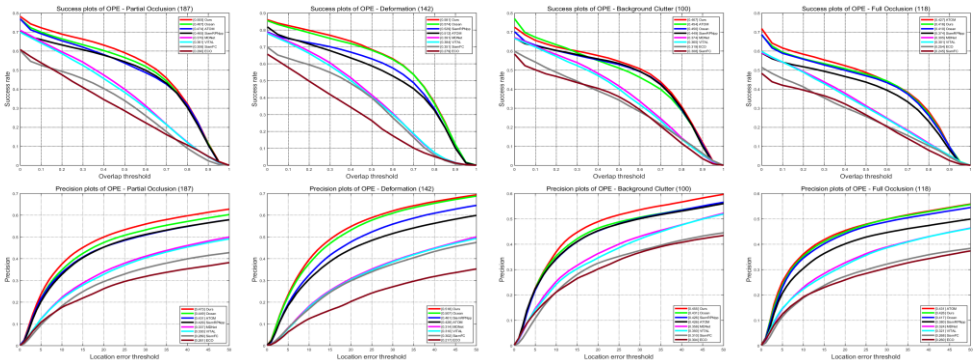

Fig 8: Success plot and Precision plot for the four disturbance conditions in LaSOT

V. Conclusion

In this paper, the RFB module is introduced to strengthen the template features and search area features that are only represented by single scale to make them more robust to the change of target size and shape. Then this paper proposes an efficient dual attention module, which can be increased under the circumstances of less amount of calculation to efficiently capture space dimension and the channel dimension, by embedding the module Ocean branch of feature fusion algorithm of target classification can effectively capture the distance information, and highlight the target, said in response to figure in order to solve the Ocean algorithm is vulnerable to similar target jamming and the influence of interference factors such as target from vision. Full experiments on two data sets, OTB2015 and LaSOT, show that the proposed algorithm has a greater improvement in tracking success rate and precision compared with the baseline algorithm Ocean, and the running speed can still be maintained at 43FPS. Compared with other mainstream tracking algorithms, the proposed algorithm also achieves leading performance.

References

[1] M.Z. Seyed Mojtaba, C. Li, G.Y. Hossein, et al., "Deep learning for visual tracking: A comprehensive survey," IEEE Transactions on Intelligent Transportation Systems, vol. 1, pp. 1-26, 2021.

[2] K.H. Zhang, L. Zhang, M.H. Yang, "Fast compressive tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, pp. 2002-2015, 2014.

[3] L. Bertinetto, J. Valmadre, J.F. Henriques, et al., "Fully-convolutional siamese networks for object tracking," Proceedings of the European Conference on Computer Vision, vol. 1, pp. 850-865, 2016.

[4] Z.P. Zhang, H.W. Peng, "Deeper and wider siamese networks for real-time visual tracking," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 4591-4600, 2019.

[5] B. Li, J.J. Yan, W. Wu, et al., "High performance visual tracking with siamese region proposal network," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 8971-8980, 2018.

[6] Z.P. Zhang, H.W. Peng, J.L. Fu, et al., "Ocean: Object-aware anchor-free tracking," Proceedings of the European Conference on Computer Vision. 1:771-787, 2020.

[7] X.L. Wang, R. Girshick, A. Gupta, et al., "Non-local neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 7794-7803, 2018.

[8] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 7132-7141, 2018.

[9] S. Woo, J. Park, J.Y. Lee, et al., "Cbam: Convolutional block attention module," Proceedings of the European Conference on Computer Vision, vol. 1, pp. 3-19, 2018.

[10] J. Fu, J. Liu, H.J. Tian, et al., "Dual attention network for scene segmentation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 3146-3154, 2019.

[11] Y. Cao, J.R. Xu, S. Lin,et al., "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops 1: 1-10, 2019.
[12] P. Zhang, Z.F. Wang, "Learning non-local representation for visual tracking," Chinese Conference on Pattern Recognition and Computer Vision, vol. 1, pp. 209-220, 2018.
[13] Q.S. Liu, J.Q. Fan, H.H. Song, et al., "Visual tracking via nonlocal similarity learning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, pp. 2826-2835, 2017.
[14] M.H. Abdelpakey, M.S. Shehata, M.M. Mohamed, "Denssiam: End-to-end densely-siamese network with self-attention model for object tracking," International Symposium on Visual Computing, vol. 1, pp. 463-473, 2018.
[15] S.T. Liu, D. Huang, "Receptive field block net for accurate and fast object detection," Proceedings of the European Conference on Computer Vision, vol. 1, pp. 385-400, 2018.
[16] Y. Wu, J.W. Lim, M.H. Yang, "Object tracking benchmark," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 37, pp. 1834-1848, 2015.
[17] H. Fan, H.X. Bai, L.T. Lin, et al., "Lasot: A high-quality large-scale single object tracking benchmark," International Journal of Computer Vision, vol. 129, pp. 439-461, 2021.
[18] Ashish, S. Noam, P. Niki, et al., "Attention is all you need," Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 1, pp. 6000-6010, 2017.
[19] F. Wang, M.Q. Jiang, C. Qian, et al., "Residual attention network for image classification," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 3156-3164, 2017.
[20] [20] Y.D. Yang, X.F. Wang, B.W. Sun, et al., "Channel expansion convolutional network for image classification," IEEE Access, vol. 8, pp. 178414-178424, 2020.
[21] Y.H. Yuan, L. Huang, J.Y. Guo, et al., "OCNet: Object context for semantic segmentation," International Journal of Computer Vision, vol. 1, pp.1-24.
[22] Z.L. Huang, X.G. Wang, L.C. Huang, et al., "Ccnet: Criss-cross attention for semantic segmentation," Proceedings of the IEEE/CVF International Conference on Computer Vision, vol. 1, no. 603-612, 2019.
[23] H.S. Zhao, Y. Zhang, S. Liu, et al., "Psanet: Point-wise spatial attention network for scene parsing," Proceedings of the European Conference on Computer Vision, vol. 1, pp. 267-283, 2018.
[24] S. Christian, I. Sergey, V. Vincent, et al., "Inception-v4, inception-resnet and the impact of residual connections on learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, pp. 1-7, 2017.
[25] L.C. Chen, P. George, S. Florian, et al., "Rethinking atrous convolution for semantic image segmentation," Proceedings of the IEEE/CVF International Conference on Computer Vision, vol. 1, pp. 1-16, 2017.
[26] J.F. Dai, H.Z. Qi, Y.W. Xiong, et al., "Deformable convolutional networks," Proceedings of the IEEE International Conference on Computer Vision, vol. 1, pp. 764-773, 2017.
[27] K.M. He, X.Y. Zhang, S.Q. Ren, et al., "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 770-778, 2016.
[28] R. Esteban, S. Jonathon, M. Stefano, et al., "Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 5296-5305, 2017.
[29] L.H. Huang, X. Zhao, and K.Q. 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, pp. 1562-1577, 2021.
[30] T.Y. Lin, M. Michael, B. Serge, et al., "Microsoft coco: Common objects in context," European Conference on Computer Vision, vol. 1, pp. 740-755, 2014.
[31] N. Hyeonseob, B. Han, "Learning multi-domain convolutional neural networks for visual tracking," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 4293-4302, 2016.
[32] P.X. Li, B.Y. Chen, W.L. Ouyang, et al., "Gradnet: Gradient-guided network for visual object tracking," Proceedings of the IEEE/CVF International Conference on Computer Vision, vol. 1, pp. 6162-6171, 2019.

[33] B. Goutam, D. Martin, G.L. Van, et al., "Learning discriminative model prediction for tracking," Proceedings of the IEEE/CVF International Conference on Computer Vision, vol. 1, pp. 6182-6191, 2019.
[34] D. Martin, B. Goutam, Khan, et al., "Atom: Accurate tracking by overlap maximization," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 4660-4669, 2019.
[35] V. Jack, B. Luca, H. Joao, et al., "End-to-end representation learning for correlation filter based tracking," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 2805-2813, 2017.
[36] Y.B. Song, C. Ma, X.H. Wu, "Vital: Visual tracking via adversarial learning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 8990-8999, 2018.
[37] D. Martin, B. Goutam, S.K. Fahad, et al., "Eco: Efficient convolution operators for tracking," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 6638-6646, 2017.