Visual Tracking With Siamese Network Based on Fast Attention Network

LIN QIN1, YANG YANG1, (Member, IEEE), DANDAN HUANG1, NAIBO ZHU2, HAN YANG3, AND ZHISONG XU1

1School of Electronic Information Engineering, Changchun University of Science and Technology, Changchun 130013, China
2PLA Academy of Military Science, Beijing 100091, China
3Beijing Engineering Research Center of Emergency Survival Security, Beijing 102206, China

Corresponding author: Yang Yang (cloneyang@126.com)

This work was supported in part by the Department of Education of Jilin Province under Grant JJKH20211424KJ; and in part by the Department of Science and Technology of Jilin Province, China, through the Natural Science Foundation Project (20200201167JC).

ABSTRACT Visual tracking remains an open challenge, as it requires real-time and long-term accurate target prediction. Siamese network has been widely studied due to its excellent accuracy and speed. Since long-term tracking may lead to model degradation and drift, most existing algorithms cannot well solve this problem. This article proposes a new Siamese Network based on Fast Attention Network named SiamFA. This method designs an attention model, which can enhance the key and global information of the target, to obtain a more robust target model and achieve long-term tracking. At the same time, the attention model is used to obtain the potential position information of the target when calculating the similarity between the template and the search area. In addition, the attention network we design reduces many redundant operations and effectively improves computational efficiency. We utilize a multi-layer perceptron to forecast the bounding box to avoid excessive hyper-parameters. In order to verify the effectiveness of our network, we conduct tests on many commonly used datasets, such as OTB100, GOT-10k, LaSOT, TrackingNet, UAV123. Our method can achieve a success rate of 62.7% and the precision rate of 64.3% on LaSOT. At the same time, it can run at about 100fps, which exceeds the comparison network, proving that our network can run in real-time.

INDEX TERMS Visual tracking, deep learning, Siamese network, attention mechanism, multi-layer perceptron.

I. INTRODUCTION

In the past few decades, visual target tracking has received increasing attention and has always been a very active research direction [1]. We could see the wide use in different fields such as intelligent monitoring, human-computer interaction, navigation and guidance. The single target tracking task is primary in computer vision. Its main task is to track the appearance and location of the target in each subsequent frame, and determine the target at the first frame [2].

Although it has made great significance, the environment it faces is complex and changeable [3], [4], such as illumination change, occlusion, background obfuscation, target loss in the field of vision, similar objects, long-term tracking failure, etc. These factors still make tracking a challenging task.

Before the tracking task based on deep learning appears, discriminative trackers based on correlation filtering, such as DCF [5] trackers, had been widely studied. Having correlation filtering as its basis, this tracker can respond roundly and establish the model more conveniently. However, the model is based on manual design features and has poor tracking performance in complex environments. In order to improve tracking performance, ECO [6] and TFCR [7] algorithms introduce deep feature fusion into the correlation filtering. The correlation filtering algorithm based on the fusion of deep features dramatically improves the tracking accuracy and effect, but there is still a massive gap between this algorithm and the target tracking method based on deep learning.

Since the appearance of SiamFC [8], it has been widely studied for its straightforward structure, high precision and easy expansion. SiamFC network adopts AlexNet [9] network to extract features, and employs cross-correlation operation to calculate the similarity between the template and the
search area to determine the location. Meanwhile, SiamFC uses multi-scale testing to deal with scale problems, which takes plenty of time. In order to settle problems, based on the SiamFC network, SiamRPN [10] apply RPN [11] network to define many anchor boxes in advance to solve the multi-scale problem. However, this multi-scale approach requires a large number of predefined hyper-parameters. Many subsequent advanced methods are added to the SiamFC framework to improve performance, such as SiamRPN, SiamRPN++ [12], DaSiamRPN [10] and other Siamese series networks. These networks process features directly using simple cross-correlation or convolution operations to evaluate the similarity between templates and search area. When predicting the target position with this method in the next frame, it only adopts the information of the template and the current search area, but ignore the characteristics of continuous target motion. Furtherly, it costs the loss of the global information of the tracking task. At the same time, the operation of not filtering the feature vector would seriously affect the effect and accuracy of long-term tracking.

This paper constructs a Siamese network framework based on the attention mechanism, which handles the problem of long-term tracking and achieves good tracking results. First, feature vectors play a vital role in tracking tasks. We use the attention mechanism to obtain the characteristics of global information and process feature vector. On the one hand, the mechanism improves the expression ability of the model; on the other hand, it highlights important information to improve the accuracy of similarity calculation and make target position estimation more accurate. Multi-layer perceptrons inspire our framework to use a fully connected network for bounding box prediction, rather than a region proposal network.

The main contributions of our article are as follows

1) We design a high efficient attention network to improve the model’s effectiveness. The self-attention mechanism is applied to obtain the global and significant information of the output characteristics of the backbone network to enhance the long-term tracking ability of the model. The designed attention network can be adopted to regulate the target’s location to determine the similarity between the template and the search area. The multi-layer perceptron is used to predict the bounding box, eliminating a mass of hyper-parameters.

2) We simplify the structure of the attention model and reduce the redundant components. Since the computational time complexity of the attention mechanism is $O(n^2 \times d)$, we reduce the number of attention layers and improve the computational speed, and the efficiency of the model significantly.

3) Based on the designed attention network model, we proposed long-term and streamlined tracking framework, which effectively improves the accuracy and speed of tracking. Through our experimental platform, the designed algorithm is tested on OTB-100 [13], GOT-10k [14], LaSOT [15], TrackingNet [16], UAV123 [17], and can reach about 100fps.

II. RELATED WORK

Based on the framework of single target tracking, we have studied it. In this part, we summarize two most relevant parts Siamese Network and Attention Network.

A. SIAMESE NETWORK TASKS

After determining the target in the first frame, the single target-tracking task predicts the target position in the subsequent images. In Siamese Network, at the initial stage of tracking, we can determine the template region patch_z and the search area patch_x according to the ground-truth box of the first frame, and then backbone network can extract the feature vectors of the two regions, and finally, it calculates the similarity of the two vectors. Since the proposal of the SiamFC network, there has been extensive research and analysis in the academic circles because of its advantages of network simplification, high accuracy and robust scalability. Adding many advanced modules and processing methods, it has improved the tracking effect of Siamese network significantly [10]–[12], [18]. Many advanced trackers do not perform well in real-time, while SiamRPN [10] uses the RPN network to solve the multi-scale problem. Due to these improvements, traditional multi-scale testing and online fine-tuning can be abandoned, making the method run much faster. However, this method predefined many candidate anchor boxes to obtain precise target locations, resulting in a mass of hyper-parameters. We also study some algorithms for processing features, like Dsiam [19], it is a dynamic Siamese network, to learn target appearance changes and background suppression online to improve tracking performance, but it cannot deal with sudden changes in light. In addition, the over-fitting problem in Siamese network training was discovered, in RASNet [20], three attention mechanisms are proposed, which are embedded into Siamese network as a layer to further describes the contour of the target object, alleviate the overfitting problem in deep network training, and improve the discriminant ability and adaptability of the network. This method prioritizes more robust feature channels and fuses them evenly. In other ways, since the proposal of residual network, it has significantly deepened the depth of the network. ResNet [21] is used in many areas such as object detection and semantic segmentation. Though it has multiple applications in single target tracking, the tracking effect is non-ideal. Finally, SiamRPN++ discovers the fact that the pooling operation in ResNet affects the translation invariance, which leads to position deviation. As a result, the tracker learns that the positive samples are all in the center during the training process, reducing trackers’ tracking performance with deep networks. After solving this issue, studies have widely employed ResNet in this field. Through the above discussion, we design our methods from the perspectives of backbone, feature processing and elimination
of hyper-parameters. We define \( \varphi(z) \) and \( \varphi(x) \) as feature vectors extracted from ResNet for the template and search area.

**B. ATTENTION NETWORK**

Target detection tasks and target tracking tasks have high similarities. Due to the successful introduction of attention mechanism from machine translation to target detection, such as ViT [22], DERT [23] and other excellent target detection models. Inspired by these models, we adjust the corresponding self-attention and cross-attention models to adapt to the single-target tracking task and design SA self-attention and CA cross-attention modules to heighten the tracking effect. The attention mechanism is widely used in Transformer [24] architecture. This function is originally used for machine translation of NLP [25]. Because the attention module can fully mine the relevant information between sentences, it can run in parallel, and does not rely on the calculation information of the upper and lower periods, which is superior to the characteristics of RNNs [26] that serial calculation can only be carried out based on the time series of the previous frame and the last frame in machine translation. We introduce this attention mechanism into Siamese network.

SA and CA are designed to process the eigenvectors corresponding to patch\(_z\) and patch\(_x\). Input images (analogous to text information) are extracted feature vectors \( \varphi(z) \) and \( \varphi(x) \) through ResNet50. In order to facilitate calculation, we adjust the dimensions of \( \varphi(z) \) and \( \varphi(x) \) through \( 1 \times 1 \) convolution. Through the SA processing template and search area branch, the feature vectors that can better characterize the characteristics of patch\(_z\) and patch\(_x\) is selected to achieve higher precision and longer tracking.

Most models using the attention method perform better than those using convolutional neural networks only, and we take advantage of attention mechanism to deal with long-term problems. We apply the attention model to construct a single target tracking network, and replace the original cross-correlation operation by Siamese series network. We design the SA (Self-Attention) module to extract the features of the template \( A_{tem} \) and the search frame \( A_{ser} \), and find the vital information of the template and the search area through the SA module. The CA (Cross-Attention) module processes the output feature vector of the SA module, calculates the attention response value of the template in the search area, and then finds the highest response value among these output vectors in the classified regression network, so as to determine the location of the target. SiamFA employs the structure of Siamese network to increase the attention mechanism and extract the vital information of the feature vector to improve the performance of the tracker. With the excellent performance of attention mechanism in the field of target classification, DETR adopts MLP network for classification, which is composed of fully connected networks. And MLP can be used not only for classification, but also for regression, so that there is no need to predetermine the size of many regression boxes, thus avoiding adjusting too many hyper-parameters. We finally adopt MLP for classification and regression.

**III. METHODS**

In this part, we introduce our SiamFA tracker. Its structure is shown in FIGURE 1. SiamFA uses ResNet50 as the backbone network, adds attention module processing features to the original Siamese network, and adopts MLP for classification and regression. We introduce these three modules in detail below.

**A. EXTRA FEATURE**

Based on the structure of the Siamese network, it constructs a template and search area pair. The template and the search area also share the weight of the feature extraction network weight like SiamFC. When constructing the template, the first given coordinates of ground-truth captures the image, and we adjust the size of the single template patch\(_z\) to \( 3 \times 128 \times 128 \), and the size of the single search area patch\(_x\) to \( 3 \times 256 \times 256 \).
In this way, the output size of ResNet is fixed, so there is no need to change the input size of the subsequent attention network dynamically. Since the target-tracking task is global and continuous, the target position moving in the upper and lower frames would not be very much. We determine the search area, according to the four times the area of the target center point determined in the previous frame, and reduce the detection range. It can save operations while eliminating interference. ResNet50 is a series of convolution and pooling layers, so deeper features have stronger translation invariance but lack a certain degree of translation and other degeneration. Although this is more conducive to object recognition, it has lower frames would not be very much. We determine the detection range. It can save operations while eliminating interference. ResNet50 is a series of convolution and pooling layers, so deeper features have stronger translation invariance but lack a certain degree of translation and other degeneration. Although this is more conducive to object recognition, it has

$$Q, K, V$$ are the values of query, key and value respectively, and they are matrix forms of $$q, k, v$$. We divide attention mechanism into multi-head attention mechanism. Equation(2) shows that, $$W_o$$ is the weight matrix for splicing $$H_i$$.

$$Multiheadattention(Q, K, V) = Concat(H_1, \ldots, H_n)W_o$$

(2)

We divide the attention mechanism into $$n = 8$$ parts so that each head notices different features. These heads are parallel computing and can be calculated very efficiently. $$N$$ heads are spliced together to get the output of the multi-head attention mechanism, as is shown in Equation (3).

$$H_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

(3)

$$W_i^Q, W_i^K, W_i^V$$ is the weight matrix of $$Q, K, V$$. These parameters are obtained through training.

2) SA

Regarding $$\phi'(z)$$ or $$\phi'(x)$$ as the input of SA, at this time, $$Q = K = V, \phi(z)$$ and $$\phi'(x)$$ have a location attribute, which corresponds to a part of the original image. So the position encoding is added in $$Q, K$$.

$$\phi'(z)$$ or $$\phi'(x)$$ is input to multiple attention outputs to form the residual structure. After LayerNorm1 regularization, the result is $$X_1$$, as shown in Equation(4).

$$X_1 = LayerNorm1[Q + Multiheadattention(Q, K, V)]$$

(4)

Considering that the attention mechanism may not fit the complex process well, in order to enhance the expression ability of the model, two layers of linear Linear1 and Linear2 are added to form the full connection layer, and then $$X_1$$ is regularized into LayerNorm2 with residual structure, as is shown in Equation(5).

$$A_{out} = LayerNorm2[(max(0, X_1 \times W_1 + b_1)$$

$$\times W_2 + b_2) + X_1]$$

(5)

The final output of SA module output results $$A_{out}$$.  

3) CA

The structure is similar to SA. The CA structure is to find the location that best matches the template from the search area. The input of CA is the output of template branch SA as $$K$$ (key), $$V$$ (value), and the output of search region branch SA is $$Q$$ (query). That is, the information of the template branch is used as the key.

To explore the process of attention more clearly, we visualize this process. We select the image shown in FIGURE 3.
as the original image, determine the template and search area based on the branch of SiamFA.

In the right image of FIGURE 4, Red is the region with the highest weight. SA assigns different weights to different feature vectors and more to significant ones.

The right image of FIGURE 5 shows the SA output weights of the search region. Similar to the right image of FIGURE 4, different weights are assigned to the feature vectors.

Ultimately, FIGURE 6 is the CA output weights that finally determine the target’s location information. Our CA module can get information about the templates and branches of the search area.

C. MLP

The final prediction in DETR [11] are calculated by a three-layer perceptron and a linear layer with ReLU activation function and hidden layer. This structure is also called MLP. It does not require prior knowledge, NMS post-processing steps, or real anchorless, eliminating the need to adjust hyper-parameters. Through the MLP structure, we can obtain $N$ boxes by processing the CA output $A_{cross}$, from which we can select region that we are most interested in as the label, and regularize the corresponding interest coordinates. It avoids designing too many hyper-parameters and makes the model simpler.

D. TRAINING AND LOSS

MLP structure accepts feature vector and pre-background discrimination and regression. For the previous background classification and discrimination tasks, the standard cross entropy is used as Equation (6)

$$L_{class} = \frac{1}{N} \sum_i -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$  (6)

When calculating the loss value of positive samples, $y_i$ is 1, and $p_i$ is the probability of a positive sample. For the regression task, GiouLoss [27] and $L_1$ Loss are used as the loss function of the regression task, as Equation(7)

$$L_{GIOU} = 1 - IOU(b_i, b_{i,\text{truth}}) + \frac{|C - b_i \cup b_{i,\text{truth}}|}{|C|}$$  (7)

$b_i$ is the box predicted by the tracker, $b_{i,\text{truth}}$ is the box corresponding to ground-truth, and $C$ is the smallest box that can enclose $b_i$ and $b_{i,\text{truth}}$. $L_1$ Loss is calculated as Equation (8)

$$L_1 = \frac{1}{n} \sum_{i=1}^n |b_{i,\text{truth}} - b_i|$$  (8)

$b_{i,\text{truth}}$ is the value corresponding to ground-truth, and $b_i$ is the model output value. Weighting GiouLoss and $L_1$ loss calculate the loss function of the regression task.

IV. EXPERIMENTS

In this section, we carried out a detailed implementation, and compared our results with the latest methods on the five benchmark datasets OTB100 [13], GOT-10k [14], LaSOT [15], TrackingNet [16] and UAV123 [17].
TABLE 1. Compares with other trackers on LaSOT, TrackingNet, GOT-10k datasets.

| A. DETAILS OF TRACKER |  |  |
|----------------------|------------------|------------------|
| Our training model adopts GOT-10k [14], LaSOT [15], TrackingNet [16], COCO-2017 [28] as the offline training datasets. We directly read the video sequence of the corresponding dataset. Due to the limited data, we employ the conventional data expansion method to expand the data. Combined with the above design scheme, we set the size of the template area to $3 \times 128 \times 128$, and the size of the search area is $3 \times 256 \times 256$. The pre-training weights of the ResNet50 [13] network on ImageNet [29] are used as the initial weights of the feature extraction network, and the attention part is initialized by Xavier [30]. Adopting AdamW [31] as the parameter optimizer, the learning rate of the feature extraction network is set to $1e-5$, the learning rate of other parameters is set to $1e-4$, and the weight attenuation parameter is set to $1e-4$. We train on Inter(R) Core(TM) i7-10700K CPU @3.8GHz and NVIDIA RTX3090 GPU, set the batchsize to 16, and the total epochs are 800. It takes about 10 days to train the model. |  |  |
| B. ABLATION STUDY |  |  |
| In order to explore the influence of attention layer number on the tracker, we compare the effects of different attention layers on tracking and choose OTB100 [13] as the dataset |  |  |
One vs. Multi-layered layer. We first construct the model based on the six-layer attention model in Transformer [24]. FIGURE 7. shows the Accuracy and Precision curves. The accuracy can reach the highest level when the number of the layer is one. According to our observation of SA output, this may be because there are too many layers of attention, which assigns too much weight to some important areas and ignore many other details.

C. COMPARE WITH OTHER TRACKER
1) RESULTS ON GOT-10K, LaSOT, TrackingNet BENCHMARKS
In this part, we compare our performance with other good trackers on GOT-10k [14], LaSOT [15], and TrackingNet [16] datasets. TABLE 1 shows the comparison results between our model and other models. Red represents the best effect, blue represents the second best, and green is the third best.

GOT-10k. The dataset contains video clips of more than 10,000 actual moving objects and more than 1.5 million manually annotated bounding boxes. It contains a lot of challenges such as occlusions and similar objects. The results of our algorithm on this dataset are shown in TABLE 1, reaching 61.6%, 72%, and 52.5% on AO, SR0.5 and SR0.75, respectively. Although SiamFA is 2% and 3.1% worse than the best KYS [32] at AO and SR0.5, SiamFA is 0.7% better than KYS at SR0.75.

TrackigNet. The dataset uses a method to track targets using the existing large-scale target detection dataset and represents the actual scene by sampling on YouTube video, but does not annotate not every frame. Tracking network videos contain a rich distribution of target categories, which are forced to be shared between training and testing. Finally, we evaluate the performance of the tracker on an isolated test-set with similar target categories and motion.
FIGURE 9. The airplane sequence on the LaSOT dataset, a detailed comparison of the effects of SiamFA and other algorithms on long-term tracking.

distributions. SiamFA achieved 75.1853% and 70.0195% under the two indicators of Success and Precision, which are almost the same as the first SiamFC++ [34].

LaSOT. It is long-term tracking dataset. This dataset has 1400 video sequences. Each video has an average of 2512 frames, the shortest video also has 1000 frames, and the longest contains 11397 frames. It considers the visual appearance and natural language contact, marking the bounding box and adding a rich natural language description to encourage exploration and tracking, combining visual and natural language features. The video sequence of this dataset is very long, which can effectively test the effect of long-term tracking. As shown in FIGURE 8., SiamFA achieves the best results on this dataset, with a success rate and accuracy rate of 62.7% and 64.3%, respectively, 5.8% and 7.6% higher than the second place DiMP-50 [33]. On the other hand, we also test the efficiency of the model. On our platform, we test DiMP-50 can only reach about 40fps, while our SiamFA can reach about 100fps. It verifies that the attention network we design can effectively improve the long-term tracking ability and accuracy, and it also proves that our framework is sufficiently lightweight.

Our algorithm is based on the Siamese network framework. In order to verify the effect of the attention network we design on the tracking task, we respectively select SiamFC [8], SiamDW [35], SiamRPN++ [12], SiamFC++, SiamAttn [36], SiamCAR [37], SiamBAN [38] and targets of the Siamese network series, and other excellent algorithms such as MDNet [39], ECO [6], VITAL [40], GradNet [41], ATOM [42], DiMP, D3S [43], KYS, Ocean [44]. The results of the above algorithm in LaSOT, TrackingNet and GOT-10k are shown in the TABLE 1. Success and Precision are the highest on the LaSOT dataset, while Success and Precision are third and second respectively on TrackingNet, comparing with the first SiamFC++ and the second SiamAttn, there is a small gap in Success and Precision indicators. On GOT-10k, SiamFA ranks first under the indicator of SR0.75, and ranks second under AO and SR0.5. The above results show that
the attention network we design can effectively improve the tracker’s long-term tracking effect of the tracker with high accuracy.

In order to observe the tracking effect of SiamFA, we select LaSOT as the dataset to observe the tracking effect, and compare the tracking effect with SiamPRN++.

In FIGURE 9, SiamFA can accurately track the target, and the predicted regression box is close to the ground-truth. Mainly when the target is deformed in 192 frames, 1058 frames, and 3428 frames, SiamFA can still accurately track the target, reflecting the excellent robustness of SiamFA. After tracking above 4315 frames, SiamRPN++ fails to track, but our SiamFA can still track accurately until the end of the sequence, which is 7960 frames. It proves that the attention mechanism network in SiamFA performs excellently in dealing with long-term sequences.

2) RESULTS ON UAV123 DATASET

The UAV123 dataset includes 123 sequences with an average length of 915 frames. In order to verify that our model is lightweight and can achieve high accuracy, we choose to test on different datasets. Because UAV123 contains a long sequence and a tiny target, it is easy to cause model degradation. Therefore, we choose this dataset for testing, and the results are speed, success rate, and accuracy.

We select the same Siamese network trackers SiamRPN++ and SiamMask [45], and use the modern network as the backbone network. FIGURE 10. shows that SiamFA’s success rate is higher than SiamRPN++ and SiamMask’s, and its accuracy is equivalent to SiamRPN++’s but higher than SiamMask’s. When we test the UAV123 dataset, SiamFA can reach about 100 fps, while SiamMask can merely reach about 90 fps, and SiamRPN++ can only reach about 75 fps. This proves that the SiamFA has more tremendous speed advantages and higher tracking accuracy.

V. CONCLUSION

This paper proposes a single target tracking method based on the Siamese network. This method uses the attention mechanism to extract essential information and construct a simplified tracking network, which significantly improves the performance of the tracker, especially the attention mechanism for long-term tracking tasks. Through flipping and experiments, we can determine the fact that the attention mechanism can improve the performance of global tracking. At the same time, through the analysis of the experimental results, the attention mechanism significantly improves the tracking accuracy.

REFERENCES

[1] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, “Visual tracking: An experimental survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1442–1468, Jul. 2014.
[2] W. Kang, X. Li, S. Li, and G. Liu, “Corrected continuous correlation filter for long-term tracking,” IEEE Access, vol. 6, pp. 11959–11969, 2018.
[3] Y. Wu, J. Lim, and M.-H. Yang, “Online object tracking: A benchmark,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Sep. 2013, pp. 2411–2418.
[4] E. Wang, D. Wang, Y. Huang, G. Tong, S. Xu, and T. Pang, “Siamese attentional cascade keypoints network for visual object tracking,” IEEE Access, vol. 9, pp. 7243–7254, 2021.
[5] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583–596, Mar. 2015.
[6] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “ECO: Efficient convolution operators for tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6931–6939.
[7] D. Yuan, N. Fan, and Z. He, “Learning target-focusing convolutional regression model for visual object tracking,” Knowl.-Based Syst., vol. 194, Apr. 2020, Art. no. 105526.
[8] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-convolutional Siamese networks for object tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 850–865.
[9] I. K. Sutskever and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural. Inf. Process. Syst. (NIPS), vol. 25, 2012, pp. 1097–1105.

[10] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, “Distractor-aware Siamese networks for visual object tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 101–117.

[11] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[12] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, “SiamRPN++: Evolution of Siamese visual tracking with very deep networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4282–4291.

[13] Y. Wu, J. Lim, and M. H. Yang, “Object tracking benchmark,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[14] L. Huang, X. Zhao, and K. Huang, “GOT-10k: A large high-performance dataset for generic object tracking in the wild,” 2018, arXiv:1810.11984.

[15] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, “LaSOT: A high-quality benchmark for large-scale single object tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5374–5383.

[16] M. Müller, A. Bibi, S. Giancola, A. Al-Subaihi, and B. Ghanem, “Track-Net: A large-scale dataset and benchmark for object tracking in the wild,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 445–461.

[17] M. Mueller, N. Smith, and B. Ghanem, “A benchmark and simulator for UAV tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 8971–8990.

[18] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, “Learning dynamic Siamese network for visual object tracking,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1781–1789.

[19] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with Siamese region proposal network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Apr. 2018, pp. 8971–8980.

[20] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, “Learning attentions: Residual attentional Siamese network for high performance online visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 4854–4863.

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Apr. 2016, pp. 770–778.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), Sep. 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Aug. 2020, pp. 213–229.

[24] V. A. M. A. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural. Inf. Process. Syst. (NIPS), 2017, pp. 5998–6008.

[25] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and M.-H. Yang, “VITAL: Visual tracking via adversarial learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 1263–1272.

[26] Y. D. Xu, Z. Y. Wang, X. Li, Y. Yuan, and G. Yu, “SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines,” in Proc. AAAI Conf. Artifi. Intell., 2020, pp. 12549–12556.

[27] Z. Zhang and H. Peng, “Deeper and wider Siamese networks for real-time visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6728–6737.

[28] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, “SiamCAR: Siamese fully convolutional classification and regression for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 6268–6276.

[29] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, “Siamese box adaptive network for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 6667–6676.

[30] H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4293–4302.

[31] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zao, C. Shen, R. W. H. Lau, and M.-H. Yang, “VITAL: Visual tracking via adversarial learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8990–8999.

[32] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, “GradNet: Gradient-guided network for visual object tracking,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6612–6711.

[33] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “ATOM: Accurate tracking by overlap maximization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 4565–4566.

[34] A. Lukezic, J. Matas, and M. Kristan, “3DS—A discriminative single shot segmentation tracker,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 7133–7142.

[35] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, “Ocean: Object-aware anchor-free tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 771–787.

[36] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, “Fast online object tracking and segmentation: A unifying approach,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1328–1338.

LIN QIN received the bachelor's degree from the Hubei University of Automotive Technology, Hubei, China, in 2019. He is currently pursuing the master’s degree with the Changchun University of Science and Technology, Changchun, China.

His research interests include data mining, visual tracking, image processing, and computer vision.

YANG YANG (Member, IEEE) received the M.S. and Ph.D. degrees from the Changchun University of Science and Technology, Changchun, China, in 2005 and 2013, respectively. He was a Visiting Scholar with the University of Notre Dame, USA, in 2011. He is currently a Professor with the Changchun University of Science and Technology. His research interests include cyber physical systems, intelligent algorithms, and data-driven control.
DANDAN HUANG received the M.S. degree from Northeastern University, Shenyang, China, in 2009, and the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 2016. Currently, she is a Lecturer with the Changchun University of Science and Technology, Changchun, China. Her research interests include image processing and machine vision.

NAIBO ZHU received the M.S. degree from the Space Tests and Launch Engineering, Equipment Command and Technical College, Beijing, China, in 1999, and the Ph.D. degree from the Space Combat and Command Equipment College, Beijing, in 2006. He is a Professor with the Department of Systems Engineering, PLA Academy of Military Science, China. His research interests include equipment test, comprehensive demonstration of equipment, and combat test and military in-service evaluation.

HAN YANG is working with the Beijing Engineering Research Center of Emergency Survival Security, China. His research interests include the Internet of Things and emergencies.

ZHISONG XU received the bachelor’s degree from the Changchun University of Science and Technology, where he is currently pursing the master’s degree. His research interests include artificial intelligence, fractional calculus, and fractional control.

---

**IEEE Access**

L. Qin et al.: Visual Tracking With Siamese Network Based on Fast Attention Network

---

VOLUME 10, 2022