A method of real-temporal object tracking combined the temporal information and spatial information

Xiaoshuo Jia1 · Zhihui Li1 · Kangshun Li1,2 · Shangyou Zeng3

Abstract
The purpose of single target tracking is to accurately and continuously locate a specific object when it is moving. However, when the objects encounter with fast movement, severe occlusion, too small size, and the same local features, the tracking algorithm which based on correlation filter or convolutional neural network will appear the positioning error phenomenon. Aiming at the above problems, this paper designs a single target tracking algorithm: relative temporal spatial network (RTSnet). RTSnet is a multi-thread network that composed of Relative temporal Information Network (RTInet) and Relative Spatial Information Network (RSInet). RTInet is designed on the basis of LSTM, and it has the predictable characteristics of temporal. It mainly obtains the relative temporal information between the frames before and after the target. RSInet, an improved twin network based on the Triplet Network, has the effect of similarity determination which can to obtain the spatial information between the frames before and after the target. In the experiments, the RTSnet is trained by using LASOT data set and verified by using the LASOT test set and the OTB100 data set. In the test set of LASOT, the accuracy of RTSnet reaches 85.5%, Trans-T reaches 62.3% and STMTrack reaches 57.4%. Meanwhile, its tracking speed reaches 117.3fps due to the RTSnet adopts dual-thread operation. On the OTB100 data-set, the accuracy of RTSnet is 81.1%.

Keywords
Single target tracking · Temporal information · Spatial information · LSTM · Triplet network

1 Introduction
Target tracking is an accurate and real-time positioning process for continuously moving targets. This technology has been widely used in the field of computer vision, such as video surveillance, robots, and drone positioning. According to the number of targets, target tracking can be divided into single target tracking and multiple target tracking. The single target tracking is the most basic theoretical model in target tracking. That model mainly includes tracking algorithms based on correlation filtering (CF) (Henriques et al. 2012; Li & Zhu 2014; Galoogahi et al. 2017; Wang et al. 2017; Possegger et al. 2015) and tracking algorithms based on the convolutional neural networks (CNN) algorithms (Ma et al. 2018; Sandler et al. 2018; Iandola et al. 2016; Szegedy et al. 2015; Simonyan and Zisserman 2014). This paper proposes a new single-target tracking algorithm from the perspective of relative information generated in the continuous movement of the target.

The single target tracking algorithm based on CF mainly uses the filter template to perform correlation algorithm processing on each frame of the input image and achieve the continuous positioning effect of the target. Among them, these algorithms include MOSSE (Bolme et al. 2010), KCF (Henriques et al. 2015), SRRDCF (Danelljan et al. 2015), DSST (Danelljan et al. 2014), etc. These algorithms have certain limitations in obtaining image features. When the background of the moving target is too complex, the accuracy of image feature acquisition by this...
type of algorithm will be greatly reduced. Since 2012, experts in this field have proposed many single target tracking algorithms based on CNN, such as SiamFC (Bertinetto et al. 2016a), CFNet (Valmadre et al. 2017), SiamRPN (Li et al. 2018), and GOTURN (Held et al. 2016). Compared to tracking algorithms based on CF, the single target tracking algorithms based on CNN has achieved good results in the key indicators of tracking speed and accuracy. The single target tracking algorithms based on CNN has been greatly improved in accuracy and speed, but when the target encounter with fast motion, severe occlusion, and the same local feature, the accuracy of this type of algorithm will drop quickly. At the same time, the single target tracking algorithms based on CF cannot achieve real-time results in accuracy and speed.

In response to the phenomenon mentioned above, this paper designs a relative temporal spatial information network (RTSnet) based on the relative spatial information and the relative temporal information of the consecutive moving target. RTSnet is a combination of the relative spatial network (RSInet) and the relative temporal network (RTInet). Among RTSnet, RSInet, a trilinear network designed based on Triplet Network (Hoffer and Ailon 2015), Trans-T (Chen et al. 2021) and Learn-F (Cheng et al. 2021), mainly make use of the relative spatial information among the targets of three consecutive to predict the spatial information of the next target. RTInet is a time prediction network based on LSTM algorithm (Yang et al. 2016; Zhou and Xu 2015; Greve et al. 2016; Gulcehre et al. 2016) and STMTrack (Fu et al. 2021), which can predict the temporal information of the next target through the relative temporal information among the targets of three consecutive. Compared with STMTrack, Trans-T and Learn-F, RTInet and RSInet will do multi-thread parallel operations. Then, RTSnet fuse the spatial information and the temporal information about the target from the RSInet and the RTInet, respectively, to obtain the target information about the next frame.

The main contributions of this paper can be summarized as follows:

1. This paper proposes an efficient tracking algorithm for multi-type information fusion. Here, the relative spatial information and relative temporal information are fused. On the one hand, the fusion of multiple information improves the algorithm’s accuracy; on the other hand, the relative processing method between consecutive frames improves the running speed of the algorithm.

2. In the extraction of spatial information, this paper designs the RSInet model. Compared with the traditional feature extraction algorithm, RSInet is designed based on the tri-linear twin network, and obtains the relative spatial information of the next frame of the target by processing three consecutive frames of images.

3. In the extraction of temporal information, this paper designs the RTInet model. Compared with the traditional temporal information feature extraction method, RTInet is improved based on the LSTM model, and obtains the relative temporal information about the next frame of the target by reading three consecutive frames of images.

4. Experiments on LASOT (Fan, et al. 2018) and OTB100 (Wu et al. 2015) data sets prove that spatial information is beneficial to target tracking. At the same time, when comparing with some SOTA trackers algorithms, it demonstrates the superiority of the algorithm in this paper.

2 Dataset and prior-work

This section mainly introduces the relevant data sets and the design ideas of RTInet and RSInet, and explains the network structure of RTSnet.

2.1 Dataset

LASOT is a long-term tracking data set. The data set has 1400 video sequences and each video has an average of 2512 frames. The shortest video in this data set also has 1000 frames, and the longest contains 11,397 frames.

The OTB100 data set contains 98 videos and 100 test scenarios. The coordinates of the object to be located in the picture are recorded in the groundtruth_rect of the data set. Each line of coordinates corresponds to the coordinate position of the upper left corner of the positioning frame and the width and height of the positioning frame in the picture.

TD: We first extract the specific location information about the target from both LASOT and OTB100. Then, we subtract the current location information from the location information of the next moment and obtain the relative time information about the target. The collection of these relative time information is the data set TD.

SD: We first crop the images of each sequence in LASOT and OTB100 according to the target position at the corresponding time, and the image set obtained after the interception is the spatial information data set of the target.

2.2 RTInet

We can draw a conclusion from StruckSiam that the position information of moving object in temporal has
certain continuous features. We will further calculate the relative temporal information of the target on the basis of that conclusion. As shown in Fig. 1, a is the target position sequence of airplane-6 from the LASOT and b is the relative position sequence of the target. It can be seen from Fig. 1 that the data fluctuation of a presents discrete features, while the b presents a relatively stable effect. Therefore, we can obtain more accurate the relative temporal feature information about the target through the calculation method of b.

On the basis of the above conclusion, we design an RTInet algorithm which can extract the relative temporal information, as depicted in Fig. 2.

RTInet is designed on the basis of the LSTM algorithm and trained on the TD dataset. The data in the TD dataset is obtained by anno subtract anno-1, which anno and anno-1 are the temporal information about the target at the current moment and the previous moment, respectively, as shown in formula 1. We will get the $\gamma_t\{x, y\}$ by the center point (x, y) of the target position from each frame and do the subtraction between the current frame and the previous frame. The data of $\gamma_t\{x, y\}$ indicates in which direction the target position of the next frame will be located. Then, we put the absolute value of $r_t$ and $r_{t-1}$ into RTInet to calculate the predicted value $r_{t+1}$ of the next frame. $r_{t+1}$ is an absolute value data. When $r_{t+1}$ is added to anno, we will get prediction data about the next frame in 16 different directions. In order to obtain more accurate prediction data, we will further to filter these data according to the angle of $\gamma_t\{x, y\}$ and the relative spatial information.

$$r_t = anno_t - anno_{t-1}$$

$$\gamma_t\{x, y\} = \frac{ano_{t-1}\{x, y\} - anno_{t-1}\{x, y\}}{anno_{t-1}\{x, y\}}$$

$$r_{t+1} = RTInet(r_{t-1}, r_t)$$

RTInet first obtains the target’s temporal information from three consecutive frames $P_{t-2}$, $P_{t-1}$, and $P_t$, namely the target’s position information anno$_{t-2}$, anno$_{t-1}$ and anno$_t$. Then, the relative temporal information $r_{t-1}, r_t$ is obtained by formula 1. Finally, the relative temporal information $r_{t-1}$ and $r_t$ are calculated by rti to obtain the relative temporal information $r_{t+1}$ of the next frame. RTInet mainly predicts the relative temporal information of the target in the next frame from the relative temporal information of three consecutive frames. Compared with STMTrack, RTInet mainly predicts changes in relative time information, reducing the error rate.

2.3 RSInet

The moving target not only has temporal information, but also has spatial information. For example, SiamFC and SiamRFC use the spatial information of the target to achieve continuous positioning of the target. When we analyze the moving target, we will divide each image into foreground and background to analyze separately. The foreground mainly describes the spatial information of the target, and the background mainly describes the background information of the target. If the spatial information about the target is obtained from the entire image, it will be disturbed by the background information so that will not only affect the speed of calculation, but also affect the accuracy of the target information acquisition. As for these problems, we propose relative spatial information and design a relative spatial information network, RSInet.

As shown in Fig. 3, we first crop three corresponding target images from three consecutive moving images, so that we can only focus on the spatial information of the target and avoid the interference of background information r_{t+1} of the next frame. RTInet mainly predicts the relative temporal information of the target in the next frame from the relative temporal information of three consecutive frames. Compared with STMTrack, RTInet mainly predicts changes in relative time information, reducing the error rate.
information. After the three intercepted target images are subtracted between the front and rear frames, two relative images about the target will be obtained. Here, RSInet is used to learn the relationship between consecutive relative images to predict the information of the next frame of target image and the RSInet’s process is shown in Fig. 4.

RSInet firstly extracts three target images information $c_{t-2}$, $c_{t-1}$, $c_t$ from the three consecutive frames of images $P_{t-2}$, $P_{t-1}$, $P_t$ according to the target position. Then, we put $c_{t-2}$, $c_{t-1}$, $c_t$ into rsi which is the feature extraction part of RSInet. And rsi is mainly a feature extraction layer composed of $3 \times 1$ and $1 \times 3$ types of convolutional layers. The parameters are shown in Table 1. Images $c_{t-2}$, $c_{t-1}$, $c_t$ are extracted by rsi to obtain corresponding high-dimensional features $f_{t-2}$, $f_{t-1}$, $f_t$. The high-dimensional feature $f_{t-2}$, $f_{t-1}$, $f_t$ obtain the parameter $\mu$ in the following formula 4. The parameter $\mu$ mainly represents the relative information ratio of the spatial feature information generated by the target between three consecutive frames. We can get the relative vector $\overline{r_{t+1}}(x, y)$ of the target center position in the next frame by multiplying $\overline{r_t}(x, y)$ and $\mu$.

$$\mu = \frac{\mu_t}{\mu_{t-1}}$$

RSInet mainly obtains the spatial relative information of the target from three consecutive frames and predicts the target information in the next frame. In terms of spatial feature extraction, compared with the relation detector (RD) of Learn-F and the attention mechanism of Trans-T, RSInet utilize the Triplet Network to obtain the relation spatial feature of continuous targets, which is more concerned with the relative spatial information of the target. Therefore, the interference of background noise is reduced, and the accuracy of relative feature information is improved.

2.4 RTSnet

RTInet extracts the relative temporal information about the target from three consecutive frames and predicts the relative temporal information about the target in the next frame. RSInet extracts relative spatial information about the target from three consecutive frames and predicts the relative spatial information about the target in the next frame under rsi. Here, we use multi-threaded parallel computing to fuse the two types of information, as shown in Fig. 5.

The main thread mainly reads information about the image. The RSInet thread reads the target’s spatial information $c_{t-2}$, $c_{t-1}$, $c_t$ from $P_{t-2}$, $P_{t-1}$, $P_t$ and obtains the relative spatial information ratio $\mu$ of the target in the next frame. The RTInet thread reads the relative temporal information $r_{t-1}$, $r_t$ about the target from $P_{t-2}$, $P_{t-1}$, $P_t$ and obtains the relative temporal information $r_{t+1}$ about the target in the next frame. Then, to calculate the target information $annot_{t+1}$ at the current moment, the relative temporal information $r_{t+1}$, and the relative spatial information ratio $\mu$ through the following formula to obtain the position information $annot_{t+1}$ about the target in the next frame.

3 Experiments

In this section, we discuss the details of RTSnet in training. And compared with some SOTA algorithms from two perspectives of algorithm accuracy and tracking speed under LASOT test set and OTB100 test set, respectively.

3.1 Experiments preparation

Experimental environment: The test platform of the network is AMD Ryzen 3 2200G; the training platform of the network is GeForce RTX 2070.

Training details: The purpose of RSInet is to obtain the relative spatial information of the next frame through the relative spatial information between successive frames. When we train RSInet, we first extract the target image $P_{t-1}$ and $P_t$ of two consecutive frames from each sequence and designate them as similar images, then extract the target image $P_{t+50}$ of the 50th frame after $P_t$ as non-similar images. Finally, we train RSInet by making the distance between $P_t$ and $P_{t+50}$ much greater than the distance between $P_t$ and $P_{t+1}$. When the value range of $\mu$ is between (0.25, 0.75), the accuracy of image similarity determination is the highest by experiment. RTInet generates relative temporal information about the next frame through the relative temporal information between successive frames. When we train RTInet, we first extract the temporal information of four consecutive frames from the sequence, namely the location information of the target, $annot_{t-2}$, $annot_{t-1}$, $annot_t$, $annot_{t+1}$. Then, the relative temporal information of $\gamma_{t-1}$, $\gamma_t$ and $\gamma_{t+1}$ between successive frames was obtained. Finally, RTInet was trained in the way of $\gamma_t$ and $\gamma_{t-1}$ fitting prediction $\gamma_{t+1}$.
Training: RTSnet’s algorithm is composed of RSInet and RTInet, and RTSnet’s algorithm also will be divided into two parts for training. RTInet’s algorithm is trained under TD dataset and optimized under formula 5, aiming to make the predicted value of RTInet close to the true value.

\[
L_{rti} = \sqrt{\sum_{i=4}^{4} RTInet(r_i^2 - r_{i+1}^2) / 4}
\]  

(5)

RSInet’s algorithm is trained under SD dataset and optimized under formula 6, aiming to let RSInet’s algorithm to reduce the relative distance \(\Delta f_{t-1,t-2}\) between similar image pairs and enlarge the relative distance \(\Delta f_{t,t-1}\) between non-similar image pairs.

\[
L_{rsi} = \Delta f_{t-1,t-2} - 2 \times \Delta f_{t,t-1}
\]

(6)

3.2 Experimental results on LASOT

We utilize the LASOT toolkit to evaluate the tracking effect of the RTSnet and its contrasting algorithm. The evaluation criteria mainly depend on two aspects: precision plots and success plots. The success plots is that under a threshold, we firstly calculate the IOU (Ren et al. 2017) overlap between the predicted border of each frame and the ground truth of the frame, and then compare the comparison rate between the IOU overlap and the threshold. The precision plots is that the ratios of frames where the location errors are within a certain values. We sort the evaluated trackers according to the area size under the curve scores of the success graph and the accuracy graph.

In order to verify the efficiency and accuracy of RTSnet, we compare the effect of RTSnet with some SOTA tracking networks, including STMTrack, Trans-T, Learn-F, MDNet (Nam Han 2016), Staple (Bertinetto et al. 2016b), SRDCF, CSRDCF (Lukezic et al. 2017), Struck (Hare et al. 2015), KCF, ECO (Daneljan et al. 2017), VITAL (Song et al. 2018), StruckSiam (Zhang et al. 2018), D-Siam (Guo et al. 2017), SiamFC, etc. Among them, SRCDF, CSRDCF, and KCF are all tracking algorithms based on CF. StruckSiam, D-Siam, SiamFC all use the characteristics of

| Table 1 Convolution parameters of rsi in RSInet |
|-----------------------------------------------|
| Kernel size/ stride pad/number | Kernel size/ stride pad/number |
|-----------------------------------------------|
| Net1 3*3/2/0/32 Concat Avg_pool(2*2/1/0) 3*3/1/100 Avg_pool(3*3/2/0) |
| 1*3/1/0/64 3*1/1/0/64 3*3/2/0/32 |
| Net2 3*3/2/0/16 3*3/1/0/16 3*3/2/0/32 |
| Net3 3*3/2/0/32 1*3/1/0/64 3*1/1/0/64 3*3/2/0/32 |

Fig. 5 Rtsnet algorithm structure diagram
Siamese algorithm to achieve efficient tracking effects. The experimental results are shown in Table 2.

Compared with algorithms such as Struck-Siam, D-Siam, and CSRDCF, from the perspective of accuracy, RTSNet uses a multi-type information fusion method to ensure the diversity and accuracy of target feature information. From the perspective of tracking speed, on the one hand, RTSNet uses the relative processing between consecutive frames to reduce the amount of calculation. On the other hand, multi-threaded parallel processing is used to accelerate the calculation speed of the algorithm.

Figure 6 shows the success plots on LASOT dataset. It can be seen intuitively that the proposed RTSNet has an accuracy rate of 85.5%. And, the proposed RTSNet is 50 and 54% higher than StruckSiam and STRCF, respectively. From Table 2, we can find that the RTSNet’s tracking speed reaches the optimal value of 117.3 fps on the LaSOT dataset.

3.3 Experimental results on OTB100

The OTB100 data set contains 98 video sequences and 11 attributes. Here, we use the OTB100 toolkit to compare the RTSNet and SiamFC, CFNet, CSRDCF. And, the result is shown in Fig. 7 and Table 3. It can be concluded that RTSNet has an accuracy rate of 81.1% higher than other tracking algorithms, 21% higher than SiamFC and 19% higher than CFNet. RTSNet’s tracking speed reaches the optimal value of 102.3 fps on the OTB100 dataset.

Compared with other algorithms, RTSNet uses a multi-information fusion method to achieve an accuracy rate of 81.1% and RTSNet uses a continuous frame relative processing method and a multi-threaded parallel processing method to achieve a speed of 102.3 fps.

Table 2 The accuracy of RTSNet and some sota tracking networks under the lasot dataset

| Network      | AUC (%) | Speed (fps) |
|--------------|---------|-------------|
| RTSNet       | **0.855** | **117.3**   |
| Trans-T      | 0.623   | 50.2        |
| STMTrack     | 0.574   | 37.4        |
| Learn-F      | 0.516   | –           |
| Struck-Siam  | 0.358   | 50.2        |
| D-Siam       | 0.353   | 48.7        |
| CSR-DCF      | 0.263   | 30.1        |
| Struck       | 0.243   | –           |

Table 3 The accuracy of RTSNet and some sota tracking networks in the OTB100 dataset

| Networks      | AUC (%) | Speed (fps) |
|---------------|---------|-------------|
| RTSNet        | **0.811** | **102.3**   |
| Trans-T       | 0.642   | 48.8        |
| STMTrack      | 0.676   | 38.2        |
| Learn-F       | 0.653   | –           |
| CFNet         | 0.621   | –           |
| CSR-DCF       | 0.569   | 10.2        |
| SRDCF         | 0.440   | 6.3         |
| SiamFC        | 0.609   | 33.4        |

Bold values are represent the best values in the comparison results.
4 Discussion

In this section, we compare RTSnet with some SOTA algorithms under the 14 attributes of LASOT dataset, and make a detailed discussion on the actual tracking effect.

4.1 Attribute evaluation

All sequences in the LASOT dataset contain a total of 14 different attributes. Here, we first analyze the superiority of the dataset attributes. Here, we use the accuracy in Fig. 6 as the benchmark, and then calculate the relative changes in the accuracy of the three tracking algorithms RTSnet, StruckSiam and STRCF under 14 attributes as shown in Fig. 8.

From the accuracy changes in Fig. 8, we find that under the attributes of FM, FOC, LR, and OV, the three algorithms all have achieved a relatively large gap, but the change value of RTSnet is lower than the other two algorithms. That is mainly because of RTSnet not only adds relative spatial information in the tracking process, but also adds relative temporal information, while the StruckSiam and STRCF algorithms only add spatial information in the tracking process. In the comprehensiveness of information, RTSnet is superior to the other two algorithms. Under the MB attribute, the accuracy of the three algorithms is reduced due to the movement of the lens. Although the accuracy of the three algorithms has been lowered under the POC attribute, RTSnet can still perform high-accuracy real-temporal tracking mainly, that’s because of the temporal feature information of the target is added to RTSnet.

Then, we continue to analyze the success plots of RTSnet and the comparison network under the LASOT data set which are shown in Fig. 9. It can be intuitively concluded from Fig. 9 that RTSnet has achieved higher accuracy in all attributes. The main reason is that compared with traditional algorithms, RTSnet performs multiple types of information fusion in information acquisition.

4.2 Qualitative evaluation

Here’s a further discussion of the tracking effects of RTSnet. We selected 4 data sets of video sequences from LASOT, namely Basketball-6, Bear-6, Boat-6, Bottle-6.

As shown in Fig. 10, the RTSnet algorithm is compared with StruckSiam and STRCF algorithms in the tracking effect. In the Basketball-6 video sequence, it is intuitively found that StruckSiam has severe target loss due to long-term target tracking. In the case of high-speed ball movement and complete occlusion, STRCF also has short-term target loss. Compared with Groundtruth, RTSnet can achieve short-term prediction and tracking of the ball’s motion, and is not affected by obstructions and fast motivation. In the Bear-6 video sequence, StuckSiam will cause target tracking errors due to too many targets with the same characteristic. RTSnet and STRCF can still track the target in real temporal with the case of camera shake. In the Boat-6 video sequence, we can intuitively see that because of the target size is too small, StruckSiam and STRCF can’t achieve precise positioning of the target. On the contrary, RTSnet can realize real-temporal tracking on small objects, which based on the characteristic relationship between the front and rear frames. There are 4 identical bottles in the video sequence of Bottle-6. When one bottle is falling down and is partially occluded by another bottle, StuckSiam and STRCF may have target positioning errors due to the similar characteristics of the bottle image. On the
Fig. 9 Under the 14 attributes of lasot, the accuracy comparison chart of rtsnet and some algorithms
contrary, RTSnet can achieve the tracking effect of the bottle according to the change of the spatial feature information of the target’s previous and subsequent frames and the change of the temporal feature information between the previous and subsequent frames.

5 Conclusion

In this paper, we propose a single-target tracking network RTSnet that integrates temporal information and spatial information. RTSnet is composed of the spatial relative network RSInet and the temporal relative network RTInet. First, we analyze the relative information changes of continuous moving targets from the spatial and temporal perspectives. After that, the relative spatial information is obtained from RSInet, and the relative temporal information is obtained from RTInet. Finally, the two kinds of relative information are merged and the prediction information about the next frame of the target is obtained. Our algorithm effectively improves the accuracy and robustness of tracking by acquiring temporal and spatial information of moving targets. Therefore, the proposed tracker is higher robust to DEF, FM, FOC, and SV. As a result, the performance of RTSnet in the LASOT data set and OTB100 data set is superior to some SOTA trackers. Meanwhile, RTSnet uses the relative processing method of continuous frames and the method of multi-threaded parallel processing so that the tracking speed on the LASOT data set reaches 117.3fps.

Acknowledgements This work is supported by The Natural Science Foundation of Guangdong Province under the Grant No.2020A1515010784, The National Natural Science Foundation of China under the Grant No.61976063 and Natural Science Program of Guangdong University of Science and Technology under the Grant No. GKY-2021KYQNK-2.

Author contributions XJ conceived the algorithms, conducted experimental demonstrations, and wrote the paper; ZL wrote the paper; KL wrote the paper; SZ wrote the paper.

Funding The work was supported by the Natural Science Foundation of Guangdong Province (No. 2020A1515010784).

Data availability This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s the Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s the Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.
References

Bertinetto L, Valmadre J, Henriques J F, et al (2016a) Fully-convolutional siamese networks for object tracking. European conference on computer vision, pp 850-865. https://doi.org/10.1007/978-3-319-48881-3_56

Bertinetto L, Valmadre J, Golodetz S, et al (2016b) Staple: complementary learners for real-temporal tracking[C]. Computer vision and pattern recognition, pp 1401–1409

Bolme DS, Beveridge JR, Draper BA et al (2010) Visual object tracking using adaptive correlation filters[C]. Computer vision and pattern recognition, pp 2544–2550

Chen X, Yan B, Zhu J, Wang D, Lu H (2021) Transformer tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Cheng S, Zhong B, Li G, Liu X, Tang Z, Li X et al. (2021). Learning to filter: siamese relation network for robust tracking

Danelljan M, Hager G, Khan F S, et al (2014) Accurate Scale Estimation for Robust Visual Tracking. British Machine Vision Conference

Danelljan M, Hager G, Khan F S, et al (2015) Learning Spatially Regularized Correlation Filters for Visual Tracking[C]. International conference on computer vision, pp 4310–4318

Danelljan M, Bhat G, Khan F S, et al (2017) ECO: Efficient Convolution Operators for Tracking[C]. Computer vision and pattern recognition, pp 6931–6939

Fan H, Lin L, Yang F et al (2018) LaSOT: a high-quality benchmark for large-scale single object tracking. arXiv: Comp Vis Patt Recogn. https://doi.org/10.1007/s11263-020-01387-y

Fu Z, Liu Q, Fu Z, Y Wang (2021) Stmtrack: template-free visual tracking with space-time memory networks

Galoogahi H K, Fagg A, Lucey S, et al (2017) Learning Background-Aware Correlation Filters for Visual Tracking. International conference on computer vision, pp 1144–1152

Greve R, Jacobsen E J, Risi S, et al (2016) Evolving Neural Turing Machines for Reward-based Learning[C]. Genetic and evolutionary computation conference, pp 117–124.

Gulcehre C, Chandar S, Cho K, et al (2016) Dynamic neural network machine with soft and hard addressing schemes [J]

Guo Q, Feng W, Zhou C, et al (2017) Learning dynamic siamese network for visual object tracking[J]. International conference on computer vision, pp 1781–1789

Hare S, Golodetz S, Saffari A et al (2015) Struck: structured output tracking with kernels. IEEE Trans Patt Anal Mach Intell 38(10):2096–2109

Held D, Thrun S, Savarese S, et al (2016) Learning to track at 100 FPS with deep regression networks. European conference on computer vision, pp 749–765. http://dpi.org/https://doi.org/10.1007/978-3-319-46448-0_45

Henriques JF, Caseiro R, Martins P et al (2015) High-speed tracking with kernelized correlation filters. IEEE Trans Patt Anal Mach Intell 37(3):583–596

Henriques J F, Caseiro R, Martins P, et al (2012) Exploring the circulant structure of tracking-by-detection with kernels[C]. European Conference on computer vision, pp 702–715

Hoffer E, Ailon N (2015) Deep metric learning using triplet network [J]. http://doi.org/https://doi.org/10.1007/978-3-319-24261-3_7

Landola F N, Han S, Moskewicz M W, et al (2016) SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size [J]. https://doi.org/10.1007/978-1-4842-6168-2_7

Li B, Yan J, Wu W, et al (2018) High performance visual tracking with siamese region proposal network. Computer vision and pattern recognition, pp 8971–8980

Li Y, Zhu J (2014) A scale adaptive kernel correlation filter tracker with feature integration. Eur Conf Computer Vision. https://doi.org/10.1007/978-3-319-16181-5_18

Lukezie A, Vojir T, Zajc L C, et al (2017) Discriminative correlation filter with channel and spatial reliability[C]. Computer vision and pattern recognition, pp 4847–4856. https://doi.org/10.1007/s11263-017-1061-3

Ma N, Zhang X, Zheng H T, et al (2018) ShuffleNet V2: practical guidelines for efficient CNN architecture design [J]. https://doi.org/10.1007/978-3-030-01264-9_8

Nam H, Han B (2016) Learning multi-domain convolutional neural networks for visual tracking[C]. Computer vision and pattern recognition, pp 4293–4302

Possegger H, Mauthner T, Bischof H, et al (2015) In defense of color-based model-free tracking[C]. Computer vision and pattern recognition, pp 2113–2120.

Ren S, He K, Girshick RB et al (2017) Faster R-CNN: towards real-temporal object detection with region proposal networks. IEEE Trans Patt Anal Mach Intell 39(6):1137–1149

Sandler M, Howard A, Zhu M, et al (2018) Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation [J].

Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition [J].

Song Y, Ma C, Wu X, et al (2018) VITAL: visual tracking via adversarial learning[C]. Computer vision and pattern recognition, pp 8990–8999

Szegedy C, Liu W, Jia Y, et al (2015) Going deeper with convolutions[C]. Computer vision and pattern recognition, pp 1–9.

Valmadre J, Bertinetto L, Henriques J F, et al (2017) End-to-End Representation Learning for Correlation Filter Based Tracking[C]. Computer vision and pattern recognition, pp 5000–5008.

Wang M, Liu Y, Huang Z, et al (2017) Large Margin Object Tracking with Circulant Feature Maps[C]. Computer vision and pattern recognition, pp 4800–4808.

Wu Y, Lim J, Yang MH (2015) Object tracking benchmark. IEEE Trans Patt Anal Mach Intell 37(9):1834–1848

Yang DD, Cai YZ, Mao N et al (2016) Long-term object tracking using adaptive correlation filters[C]. Computer vision and pattern recognition, pp 4293–4302

Zhang Y, Wang L, Qi J, et al (2018) Structured siamese network for large-scale single object tracking. arXiv: Comp Vis Patt Recogn. https://doi.org/10.11263-020-01387-y

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.