Single object tracking algorithm in occlusion scene based on improved ECO

Guorong Xie¹a, Rongqi Jiang¹b, Yi Qu²c*

¹ Graduate Team, Engineering University of PAP, Xi’an Shanxi, China
² Information Engineering College, Engineering University of PAP, Xi’an Shanxi, China
▫email: wuxingxiedeyx@163.com, ▫email: jjqjjqq163@163.com
▫ Corresponding author’s e-mail: ▫email: wjquyi@sina.com

Abstract: To alleviate the occlusion problem in a single object tracking scene, this paper proposes an ECO-MHDU object tracking algorithm with a more powerful anti-occlusion performance based on the ECO tracker. The algorithm first uses the pre-trained MobileNetV3 lightweight backbone network on the ImageNet dataset to replace the ResNet network in the ECO to increase the speed of the algorithm to obtain the shallow and deep feature information of the image, while effectively using the attention mechanism in the MobileNetV3 network to strengthen the algorithm’s ability to extract target features; secondly, use the DropBlock operation on the acquired feature map to generate a random continuous mask on the feature map channel to improve the algorithm’s learning of the global robust spatial structure information; finally, a confidence update strategy is introduced into the GMM sample generation space. To improve the quality of training samples, unreliable tracking states such as confidence detection and occlusion are designed to avoid updating the sample space with damaging information. Compared with the ECO algorithm, the ECO-MHDU algorithm proposed in this paper has a success rate of 68.0% on the occlusion attributes of the OTB100 dataset, which is 2.3% higher than the ECO algorithm, and the ECO-MHDU algorithm also showed the best performance on the entire dataset sequence, with a success rate of 69.3%.

1. Introduction

Object tracking algorithms based on correlation filter are one of the mainstream research directions to solve tracking problems. In recent years, with the unremitting efforts of researchers, correlation filter trackers combining manual features[1, 2] have gradually developed to trackers using depth features[3]. At the same time, there is a tracking algorithm that transforms the correlation filter into a convolutional layer embedded network[4, 5], among which ECO[6] is a very representative tracking algorithm that shows strong tracking performance. An important reason for tracking failure when the target is occluded[7] is the focus and difficulty of current research. In order to better improve the tracking performance of the algorithm and alleviate the occlusion problem under the object tracking task, the researchers combined the object tracking algorithm based on correlation filter with the deep convolutional network with outstanding feature expression ability and proposed a high-reliability tracking algorithm based on the ECO object tracking algorithm. At the same time, due to the deep network structure of CNN, the network has strengthened the acquisition of high semantic information
of the target, and also improved the "drift" phenomenon of the algorithm tracking the target in the occluded scene to a certain extent.

However, the current problems of object tracking tasks in occlusion scenarios are still very serious. The reasons for the inadaptability of the tracking framework based on correlation filter to occlusion problems are:

- The online samples obtained by the cyclic intensive sampling method are highly similar, the generalization performance of the model is easily impaired, and it is difficult to distinguish the target object that is deformed due to occlusion.
- The online collection of samples is affected by the state of the tracking period, and it is easy to introduce damaged samples such as occlusion and rotation in and out of the plane, which makes the model fit the occlusion and other bad information and cause error accumulation.
- The occlusion attribute conforms to the long-tailed distribution in the training sample set, and the effective information available for model learning is insufficient; fourth, the feature extraction module architecture provides semantic information, which is still sensitive to occlusion, and lacks sufficient information to distinguish similar semantic interference.

In response to the above problems, this paper, based on the current powerful ECO tracking algorithm based on correlation filter, proposes an ECO-MHDU object tracking algorithm with more powerful anti-occlusion performance. The algorithm implements three effective anti-occlusion mechanism designs based on ECO: 1) replace the ResNet network in the ECO with the pre-trained MobileNetV3 lightweight backbone network on the ImageNet dataset to improve the speed of the algorithm to obtain the shallow and deep feature information of the image, and effectively use the attention mechanism in the MobileNetV3 network to strengthen the algorithm's ability to extract target features; 2) using the DropBlock operation to generate random continuous masks on the feature map channels extracted by the CNN network, generate hard positive samples of class occlusion, forcing the model to learn global robust spatial structure information when facing damage to the target due to occlusion, and to learn how to recover damaged samples; 3) the confidence update strategy is introduced into the GMM sample generation space. In order to improve the quality of training samples, unreliable tracking states such as confidence detection and occlusion are designed to avoid updating the sample space with damaged information.

2. ECO correlation filter tracking algorithm

The algorithm ECO is based on the correlation filter tracker C-COT [3], and analyzes three important factors that affect the tracking rate, including model size, training set size, and model update strategy, and taking the improvement of time efficiency and space efficiency as the starting point, a convolution operation using factorization, a compact generation model of sample space, and an equal interval model update strategy are proposed, and Resnet50, HOG, CN are used to extract multiple features of the search area and achieve powerful tracking accuracy.

2.1 Factorized Convolution Operator

The integration of high-dimensional feature maps will lead to a sharp increase in the parameters of the tracking object appearance model, often exceeding the dimensionality of the input image. In the C-COT algorithm model, it is necessary to continuously update more than 800,000 parameters online at the same time, resulting in a greatly reduced tracking rate, and it is not suitable for tracking tasks with scarce training data, because this may cause the tracking model to overfit. The ECO algorithm addresses the above-mentioned C-COT algorithm problems and introduces the decomposition convolution factor to achieve the effect of reducing the number of model parameters and eliminating redundancy.

Usually, each dimensional feature corresponds to a training filter, but in fact, most filters have less energy, especially high-dimensional depth features that are not conducive to positioning, so there is a problem of relative redundancy. The ECO algorithm is optimized for this, choosing a filter that contributes more, and each dimensional feature can be represented by a linear combination of C filters.
The characteristic layer $d$ corresponds to the filter with a linear weighted structure of the learnable coefficient $P_{d,c}$. The new multi-channel detection function is as follows:

$$S_{pf}\{x\} = Pf \ast J \{x\} = \sum_{c,d} P_{d,c} f^c \ast J_d \{x^d\} = f \ast P^T J \{x\} \quad (1)$$

Among them, the $D \times C$ matrix $P$ compactly represents the linear combination coefficients $P_{d,c}$, which are only learned in the first frame, and the subsequent frames remain unchanged. By minimizing the classification error of the decomposition factor, and jointly learning the filter and matrix, the objective function of the filter learning and training is as follows:

$$E(f, P) = \|z^T P f - 1\|^2 + \sum_{c=1}^{C} \|\hat{w} \ast \hat{f}^c\|_2^2 + \lambda \|P\|^2 \quad (2)$$

In the formula, $z = J \{x\}$ is the interpolation feature map, and the Frobenius norm of $P$ is added to it, controlled by the parameter $\lambda$, and introduced through the decomposition factor to achieve a fundamental reduction in the amount of data. Finally, the conjugate iteration method of Gauss-Newton and Conjugate Gradient is used to optimize the filter.

### 2.2 Generative Sample Space Model

The tracking algorithm based on the correlation filter realizes tracking in the way of online sampling, detection, and training. The iterative optimization process of model training and learning relies on the storage of training sample sets collected online. However, the memory is limited, especially the high-dimensional information representation model, which will also lead to greater time consumption. Generally, the typical way to maintain feasible memory consumption is to discard historical samples for a long time, but this is not friendly to target objects that change complexly during tracking. It may cause the model to overfit the recent frame mutation samples, accumulate errors, and cause tracking drift.

ECO builds a compact generation model of the sample space, adding a new sample every update frame, using Gaussian Mixture Model (GMM) to generate different groups, the samples in the group are similar, but the differences between the different groups are large. Therefore, on the basis of reducing the number of training samples, sample diversity is maintained.

Based on the joint probability distribution $p(x,y)$ of the sample feature mapping and the corresponding expected output score, the filter that minimizes the expected correlation error is trained, so the objective function is expressed as:

$$E(f) = \sum_{l=1}^{L} \|S_{f} \{\mu_l\} - y\|_2^2 + \sum_{d=1}^{D} \|w_f^d\|_2^2 \quad (3)$$

When a new sample $x_j$ appears, first initialize the parameters $\pi_m = \gamma$, $\mu_m = x_j$ of the new sample set component $m$. If the number of samples in the component exceeds the limit $L$, the component whose weight $\pi$ is less than the set threshold is discarded; otherwise, the two components $k$ and $l$ with the smallest distance are merged into a common component as shown below:

$$\mu_n = \frac{\pi_k \mu_k + \pi_l \mu_l}{\pi_k + \pi_l} \quad (4)$$

Gaussian mixture modeling through sample space reduces the size of the training set while maintaining sample diversity, alleviating the problem of similar frames in the model overfitting.

### 2.3 Model Update Strategy

Most tracking algorithms based on the correlation filter use an online continuous learning strategy, and the target template is updated every frame, but they are sensitive to occlusion, rotation in and out of the plane, and sample deformation damage such as scale changes. Like the default sample discarding method, there is a problem of over-learning to adapt to the latest frame, which at the same time leads to a decrease in the robustness and real-time performance of tracking.
Set the model update strategy of the fixed frame \( N_s \), combined with the update of each frame of the sample set, can capture the sample change information in the interval frame. The parameter setting is also more critical, if it is too large, it is easy to lose track of the target's changing appearance speed, if it is too small, it may drift due to over-fitting highly similar information and sudden error frame information. Finally, the algorithm sets the parameter \( N_s \) to 6. Although its strategy is simple, it effectively improves the tracking speed and robustness at the same time.

3. Implementation of ECO-MHDU Anti-occlusion Object Tracking Algorithm

3.1 Optimization of lightweight feature extraction module based on MobileNetV3 architecture

The ECO tracking algorithm uses multiple features of Resnet50, HOG, and CN to extract the features of the search area to achieve powerful tracking accuracy. However, the Resnet50 feature extraction network has a deeper network structure, so the algorithm will consume more time when extracting feature information through the Resnet50 network. In this regard, this paper uses the powerful and more efficient convolutional neural network MobileNetV3 [8] as the feature extraction network of the ECO-MHDU algorithm to improve the speed of the algorithm to obtain the shallow and deep feature information of the image. And effectively use the attention mechanism in the MobileNetV3 network to strengthen the algorithm's ability to extract target features.

MobileNetV3 divides the model into two versions, large and small according to actual needs. Among them, MobileNetV3-large improves the accuracy by about 3.2% and reduces the delay by 20% compared with MobileNetV2 in the classification task of ImageNet large-scale image verification set. Since the MobileNetV3-large network can obtain more powerful semantic features, and the use of high-semantic information is one of the effective means to solve the occlusion problem, this choice is based on the MobileNetV3-large improved feature module framework.

The MobileNetV3 network is stacked with multiple Bottleneck as the core unit. It uses NAS technology based on MobileNetV1[9] and MobileNetV2[10] to obtain the best structure and parameter settings of the network. The MobileNetV3-large structure is shown in Figure 1.

![Figure 1 MobileNetV3-large model structure configuration diagram](image)

It can be seen from Figure 1 that the activation function used in the MobileNetV3 algorithm includes two activation functions: h-swish and ReLU. HS represents the activation function h-swish, and RE is RELU. The network takes a 224×224 size image as input and implements 5 down-sampling operations through the bneck layer. The bneck layer is the core unit of the MobileNetV3 algorithm, Bottleneck, whose structure is shown in Figure 2.
As shown in Figure 2, the Bottleneck structure first integrates the inverse residual structure design of the linear bottleneck in the MobileNetV2 network and uses a 1×1 Conv layer at the input to upgrade the channel dimension of the feature map. In the MobileNetV3-large model, the dimension is increased to the size of \( \text{exp size} \) in the figure; then through a layer of 3×3 depth separable convolutional layer, that is, the Dwise layer in the figure, it can reduce the amount of model calculation and obtain a larger receptive field; finally, a 1×1 convolutional layer is used to restore the number of channels that have been upgraded to the size when they were input to Bottleneck; at the same time, the Bottleneck structure also adds a lightweight SE channel attention module, and its structure is shown in Figure 3.

As shown in the figure above, the SE module implements the extraction of the importance of feature maps through global pooling, channel compression, and weight distribution. Among them, in the MobileNetV3 algorithm, the FC fully connected design used in the SE module is changed to use a 1×1 convolutional layer, and the HSigmoid activation function is used to realize the distribution of access weights, which effectively simplifies the calculation of the SE module.

The core idea of the SE module is to compress the global spatial features of the feature map into the descriptors of each channel and adjust the feature map according to the dependence of each channel to improve the network's ability to characterize important channel features. The SE model can improve the algorithm's discrimination of "false" target features, and for the occlusion problem in tracking tasks, adding the attention module can also effectively improve the algorithm's ability to extract features of occluded targets, and alleviate the problem of drifting to semantic occlusions or background-similar interferences in occlusion scenes.

This paper refers to the design of CNN feature extraction in the original ECO algorithm and uses the MobileNetV3-lager network model to output the shallow features extracted by the second Bottleneck (2\(^{nd}\) downsampling) of the network and the deep features extracted after the 13th Bottleneck (4\(^{th}\) downsampling) to realize the feature extraction of CNN.

### 3.2 Based on DropBlock hard positive sample generation strategy

In actual situations, the more robust the tracker is, the more stable the tracking effect will be under occlusion. Therefore, in order to obtain a tracker with stronger generalization ability and robustness, this paper proposes to use the DropBlock [11] operation on the feature map obtained in CNN to achieve the improvement of the generalization ability of the algorithm.
DropBlock was proposed by Golnaz Ghiasi in 2018. It can effectively solve the problem of over-parameterization of deep neural networks and is suitable for situations where there is a lot of noise during training. DropBlock is a continuous area unit of randomly lost feature maps. By destroying the continuous area, it can effectively lose the semantic information of the target. During the training optimization process, the model can be forced to learn the remaining effective other types of information, thereby improving the robustness. Compared with the Dropout [12] operation, DropBlock can implement a method based on structure loss, which can effectively eliminate part of the semantic information in the feature map, making model classification more difficult, thereby improving the robustness of model training, so DropBlock is more suitable for feature maps. The following figure shows a schematic diagram of Dropout and DropBlock operations. Figure 4 (a) is the CNN input image, and the green area in Figure 4 (b) (c) contains the input image semantic information activation unit.

Figure 4 Schematic diagram of Dropout and DropBlock operations

In the experiment of this paper, refer to the paper [11] and set the input parameter keep_prob of the dropblock algorithm to 0.9 and the parameter block_size to 7.

3.3 GMM sample space generation model based on confidence update

Inspired by the algorithms MOSSE [13] and LMCF [14], this paper introduces a confidence update strategy, mainly for the purpose of detecting the occlusion state in a timely and effective manner, and designing a set of complementary effective confidence indicators to accurately determine whether the tracking state is reliable. Considering that on the basis of sample space modeling, the confidence level is used as a way to control the timing of updating and judging the target template, and it cannot avoid learning from occlusion-damaged samples. Therefore, this paper designs a confidence update strategy for the GMM sample space, combined with the target template fixed frame update method, to improve the quality of training samples and enhance the tracking robustness in occluded scenes.

The general depth features are derived from CNN, which can provide insufficient information to distinguish between classes. Therefore, when the target encounters occlusion, the model may not be able to distinguish similar interference such as occlusion objects, which is intuitively reflected in the multi-peak situation in the response graph. If the model can accurately detect the unreliable tracking state such as occlusion in time, it can resume tracking after occlusion, otherwise, it will drift to the occluded object. Regarding the occlusion attribute, it is not reliable to judge the peak height of the response graph alone. Therefore, this paper designs the following complementary confidence index values to ensure that the set of confidence indicators can change significantly in the occlusion scene. Its characteristics are as follows:

The highest response score \( F_{\max} \): the highest peak response score on the response graph \( F(s,y;w) \), which can reflect the deterministic index of the tracking effect, where \( s \) is the image block based on the target position of the previous frame, \( y = f(s;w) \) is the target position, and \( w \) is the filter template:

\[
F_{\max} = \max F(s,y;w)
\] (5)

Peak-to-Sidelobe Ratio (PSR): can quantify the definition of the correlation peak, and decompose the correlation output into peak \( F_{\max} \) and side lobes. The side lobes are pixels outside the \( 11 \times 11 \) window near the peak, \( \mu_{s} \) and \( \alpha_{s} \) respectively represent the average and standard deviation of the side lobes, which are defined as:
\[
PSR = \frac{F_{\text{max}} - \mu_s}{\alpha_s}
\]

(6)

It is generally believed that when the PSR value is in the interval of [20, 60], the peak value is very strong, and the tracking state is reliable at this time; when it is in the left and right interval, it may face occlusion or tracking failure.

Average peak-to-correlation energy (APEC): APEC is an indicator that reflects the degree of fluctuation of the response graph and the reliability of the test results, which is defined as:

\[
APCE = \frac{\min_{w,H} (F_{w,0} - F_{\text{min}})^2}{\text{mean} \left( \sum_{n,h} (F_{w,h} - F_{\text{min}})^2 \right)}
\]

(7)

Among them, is the corresponding pixel peak. When the peak is sharper and the sidelobe and other noises are less, the response graph shows a smooth drop in the area beside the sharp peak. At this time, APCE will become larger; when the target encounters occlusion or disappears, the peak will become blurred due to the increase in uncertainty, and there may be many indistinguishable peaks around, and APCE will drop sharply at this time.

The flow of the target being occluded is as follows:

A. When the target starts to be partially occluded, the response graph may fluctuate, and the APCE and PSR values will be significantly reduced;

B. When it is completely blocked, if there is no background clutter and other attributes, it is lower at this time, APCE, PSR rise slightly, but still lower than the value in A, if there is background clutter, etc., APCE, PSR will remain at a low level;

C. When the target recovers from the occlusion, the response graph fluctuates, APCE decreases, and PSR and may increase due to tracking drift.

Therefore, in this paper, the tracking that meets the conditions of A, B, and C being less than the corresponding threshold at the same time is regarded as an unreliable state such as occlusion.

This paper introduces a confidence update strategy in the sample space generation model to improve the quality and reliability of training samples. When the model detects occlusion, the GMM sample generation model is not updated to avoid model learning to fit damaged samples; at the same time, the search area is reduced, and the model learns to fit the background interference information in the case of partial information loss of the object so that after recovering from the occlusion, the target is correctly detected and tracked. The specific situation is shown in Figure 5.

As can be seen from the above figure, depending on the fluctuation of the response graph reflected by the confidence combination index, the algorithm can detect the occlusion state in time. During occlusion, the response graph has multiple peaks and fluctuates greatly. By reducing the learning rate and narrowing the search range, after the target recovers from the occlusion state, the response graph can maintain a single peak and continue to achieve robust tracking.

4. Experiment & Result Analysis
The experimental model training platform uses the hardware platform as Intel Core i7-8700K CPU, GeForce GTX 1080Ti 11G GPU; the software uses Ubuntu system, python3.7, CUDA10.2, CUDNN8.8,
mxnet1.8.0 deep learning framework. The experimental test selects the OTB100[15] dataset, and the evaluation indicators use the tracking success rate, accuracy rate, and FPS in the case of OPE.

4.1 Ablation experiment and result analysis
In this section, we will design ablation experiments to verify the tracking effectiveness of different strategies in the ECO-MHDU algorithm for occlusion and other scenes. There are 5 groups in the experiment, and the specific ideas are as follows:

- Benchmark Experiment (ECO): For the ECO performance of the algorithm, its CNN feature uses the ResNet-50 model, extracts its first down-sampling and fourth down-sampling features, and uses HOG manual features;
- Experiment 1 (ECO-D01): To verify the effectiveness of DropBlock, the algorithm is ECO-D01. In the benchmark experiment, DropBlock is applied to the training sample, setting the parameters keep_prob=0.9, block_size=7;
- Experiment 2 (ECO-U): Verify the validity of the confidence update strategy that acts on the GMM sample space. The algorithm is ECO-U. In the benchmark experiment, set the sample input confidence condition of the GMM sample space and set the parameter $f_{\text{max\_threshold}}=0.2$, $psr_{\text\_threshold}=10$, $apce_{\text\_threshold}=20$;
- Experiment 3 (ECO-MH): To verify the effectiveness of the feature extraction framework design based on MobileNetV3, in the benchmark experiment, the algorithm replaces MobileNetV3-large with the original feature extraction backbone network ResNet-50, extracts the second down-sampling and fourth down-sampling features, and merges them with the manual feature HOG;
- Experiment 4 (ECO-MHDU): Verify the ECO-MHDU algorithm proposed in this paper.

The results are shown in Figure 6 and Table 1, which is the success and precision plot of all the sequences.

![Success plot](image1)
![Precision plot](image2)

Figure 6. OTB100 dataset test results

| Tracker    | ECO   | ECO-D01 | ECO-MH  | ECO-U  | ECO-MHDU |
|------------|-------|---------|---------|--------|----------|
| Success    | 66.9% | 68.8%   | 68.4%   | 68.2%  | 69.3%    |
| Precision  | 89.0% | 90.6%   | 92.3%   | 90.7%  | 92.6%    |
| FPS        | 9.84  | 9.70    | 12.72   | 9.84   | 12.63    |

Table 1. Comparison results of ablation experiments

Combined with the results of the above ablation experiments, the three improved methods proposed in this paper can effectively improve the performance of tracker, and all have different degrees of performance improvement in the OTB100 dataset test. Among them, using the MobileNetV3 lightweight feature extraction network, the success rate has increased by 1.5%, the accuracy rate has
increased by 3.6%, and the FPS has increased by 2.88 fps; the introduction of the DropBlock strategy has increased the algorithm success rate by 1.9% and the accuracy rate by 1.6%; based on the confidence update strategy of the GMM sample generation space, the success rate has increased by 1.3%, and the accuracy rate has increased by 1.7%; while the ECO-MHDU algorithm has a success rate of 69.3% and an accuracy rate of 92.6%, and FPS has increased by 2.79 fps compared with ECO.

In order to further evaluate the effectiveness of each strategy in solving the occlusion problem, this paper compares the performance of different improved strategies under the occlusion attributes of the OTB100 dataset. The results are shown in Figure 7:

4.2 Horizontal comparative experiment and result analysis
This section designs a horizontal comparison experiment and tests on the classic OTB100 tracking dataset. A wide range of other tracking algorithms based on correlation filter are selected: ECO, manual feature HOG+CN version of ECO-HC, LDES, Staple, DSST, CSK as the benchmark model. The experimental results are shown in Figure 8.
Table 2. Horizontal comparison results of different algorithms on the OTB100 dataset

| Tacker          | CSK    | SAMF  | LDES   | Staple | ECO-HC | ECO   | ECO-MHDU |
|-----------------|--------|-------|--------|--------|--------|-------|----------|
| All sequences   | Success| 0.377 | 0.563  | 0.603  | 0.605  | 0.639 | 0.669    |
|                 | Precision| 0.495 | 0.751  | 0.778  | 0.792  | 0.852 | 0.890    |
| occlusion       | Success| 0.333 | 0.543  | 0.578  | 0.563  | 0.625 | 0.657    |
|                 | Precision| 0.414 | 0.721  | 0.743  | 0.722  | 0.838 | 0.884    |

From the above table, on the OTB100 dataset, the ECO-MHDU algorithm proposed in this paper shows the best tracking performance. At the same time, the ECO-MHDU algorithm also significantly improves the occlusion problem during tracking.

Finally, in order to better verify the effectiveness of the ECO-MHDU algorithm, this paper visually compares the video sequences of typical occlusion attributes in the test results of the ECO, ECO-HC, and ECO-MHD algorithms in the OTB100 dataset. The results are shown in Figure 9.
5. Conclusions
Aiming at the occlusion problem during tracking, this paper introduces the feature extraction backbone network optimization based on MobileNetV3 based on the ECO algorithm, the hard positive sample generation strategy based on the DropBlock class occlusion, and the GMM sample space generation model based on the confidence update. Finally, the ECO-MHDU algorithm is proposed. Through experimental tests on the OTB100 dataset, the ECO-MHDU algorithm proposed in this paper effectively improves the occlusion scene tracking problem. The success rate of the OTB100 dataset on the occlusion attribute reaches 68.0%, which is 2.3% higher than the ECO algorithm; at the same time, the ECO-MHDU algorithm proposed in this paper also shows the best performance on the entire dataset sequence, with a success rate of 69.3%. It can be seen that the algorithm proposed in this paper effectively improves the tracking performance of the ECO algorithm in occlusion scenarios, and provides more solutions for the current research on occlusion problems in tracking scenarios.

Acknowledgments
This paper is one of the phased achievements of the National Natural Science Foundation of China (No.61101238).

References
[1] Danelljan, M., Hager, G., Shahbaz Khan, F., & Felsberg, M. (2015). Learning spatially regularized correlation filters for visual tracking. In Proceedings of the IEEE international conference on computer vision. pp. 4310-4318.
[2] Danelljan, M., Häger, G., Khan, F., & Felsberg, M. (2014). Accurate scale estimation for robust visual tracking. In British Machine Vision Conference, Nottingham.
[3] Danelljan, M., Robinson, A., Khan, F. S., & Felsberg, M. (2016, October). Beyond correlation filters: Learning continuous convolution operators for visual tracking. In European conference on computer vision. Springer, Cham. pp. 472-488.
[4] Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., & Torr, P. H. (2017). End-to-end representation learning for correlation filter based tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition . pp. 2805-2813.
[5] Li, X., Ma, C., Wu, B., He, Z., & Yang, M. H. (2019). Target-aware deep tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1369-1378.

[6] Danelljan, M., Bhat, G., Shabbaz Khan, F., & Felsberg, M. (2017). Eco: Efficient convolution operators for tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6638-6646.

[7] Li, Xi, ZHa Yufei, ZHANG Tianzhu, Cui Zhen, Zuo Wangmeng, Hou Zhiqiang, Lu Huchuan, Wang Hanzi. (2019). A review of deep learning target tracking algorithms [J]. Chinese Journal of Image and Graphics. Vol. 24 No. 12: 2057-2080.

[8] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314-1324.

[9] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[10] Howard, A., Zhmoginov, A., Chen, L. C., Sandler, M., & Zhu, M. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation.

[11] Ghiasi, G., Lin, T. Y., & Le, Q. V. (2018). Dropblock: A regularization method for convolutional networks. arXiv preprint arXiv:1810.12890.

[12] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

[13] Bolme, D. S., Beveridge, J. R., Draper, B. A., & Lui, Y. M. (2010, June). Visual object tracking using adaptive correlation filters. In 2010 IEEE computer society conference on computer vision and pattern recognition (pp. 2544-2550). IEEE.

[14] Wang, M., Liu, Y., & Huang, Z. (2017). Large margin object tracking with circulant feature maps. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4021-4029.

[15] Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2411-2418.