Journal of Smart Science and Technology

PENERBIT PRESS UNIVERSITI TEKNOLOGI MARA

2023 March Volume 3 Issue 1

eISSN: 2785-924X

Performance of Correlational Filtering and Deep Learning Based Single Target Tracking Algorithms

ZhongMing Liao[1,2,*], Azlan Ismail[1,3]

[1]*School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*
[2]*XinYu College, JiangXi 338004, P.R.China*
[3]*Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Al-Khawarizmi Complex, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*

Received: 05-09-2022
Revised: 24-12-2022
Accepted: 10-01-2023
Published: 30-03-2023

*Correspondence
Email: liaozhongming168@gmail.com
(ZhongMing Liao)

DOI: https://doi.org/10.24191/jsst.v3i1.42

Abstract

Visual target tracking is an important research element in the field of computer vision. The applications are very wide. In terms of the computer vision field, deep learning has achieved remarkable results. It has broken through many complex problems that are difficult to be solved by traditional algorithms. Therefore, reviewing the visual target tracking algorithms based on deep learning from different perspectives is important. This paper closely follows the tracking framework of target tracking algorithms and discusses in detail the traditional visual target tracking methods, the mainstream single target tracking algorithms based on correlation filtering, and the video single target tracking algorithms based on deep learning. Experiments were conducted on OTB100 and VOT2018 benchmark datasets, and the experimental data obtained were analysed to derive two visual single-target tracking algorithms with optimal tracking performance. Finally, the future development of tracking algorithms is envisioned.

Keywords

Deep learning; Correlation filtering; Target tracking algorithms

Citation: Liao, Z., & Ismail, A. (2023). Performance of correlational filtering and deep learning based single target tracking algorithms. *Journal of Smart Science and Technology*, *3*(1), 63-79.

1 Introduction

Visual target tracking is a fundamental and important research topic in the field of computer vision, which has received a great deal of attention from scholars. Given the state (position and size) of a target in the first frame of a video, the aim is to predict the state of the target in subsequent frames[1,2]. Visual target tracking has wide and deep applications in human-computer interaction, intelligent video surveillance, medical diagnosis, visual navigation, and other fields.

Although visual target tracking technology has been studied for many years and some progresses have been made, it is still difficult to meet the practical needs, such as scale change, fast motion, deformation, blur, illumination change, occlusion, and background clutter in some situations. Many academics attempt to improve target tracking and overcome its problems[3,4], mainly including the challenging factors like the self-factor and background factors as shown in Figure 1. Often, multiple challenges are faced in a tracking task, which makes it particularly important to design a robust tracking algorithm that can cope with a variety of complex situations.

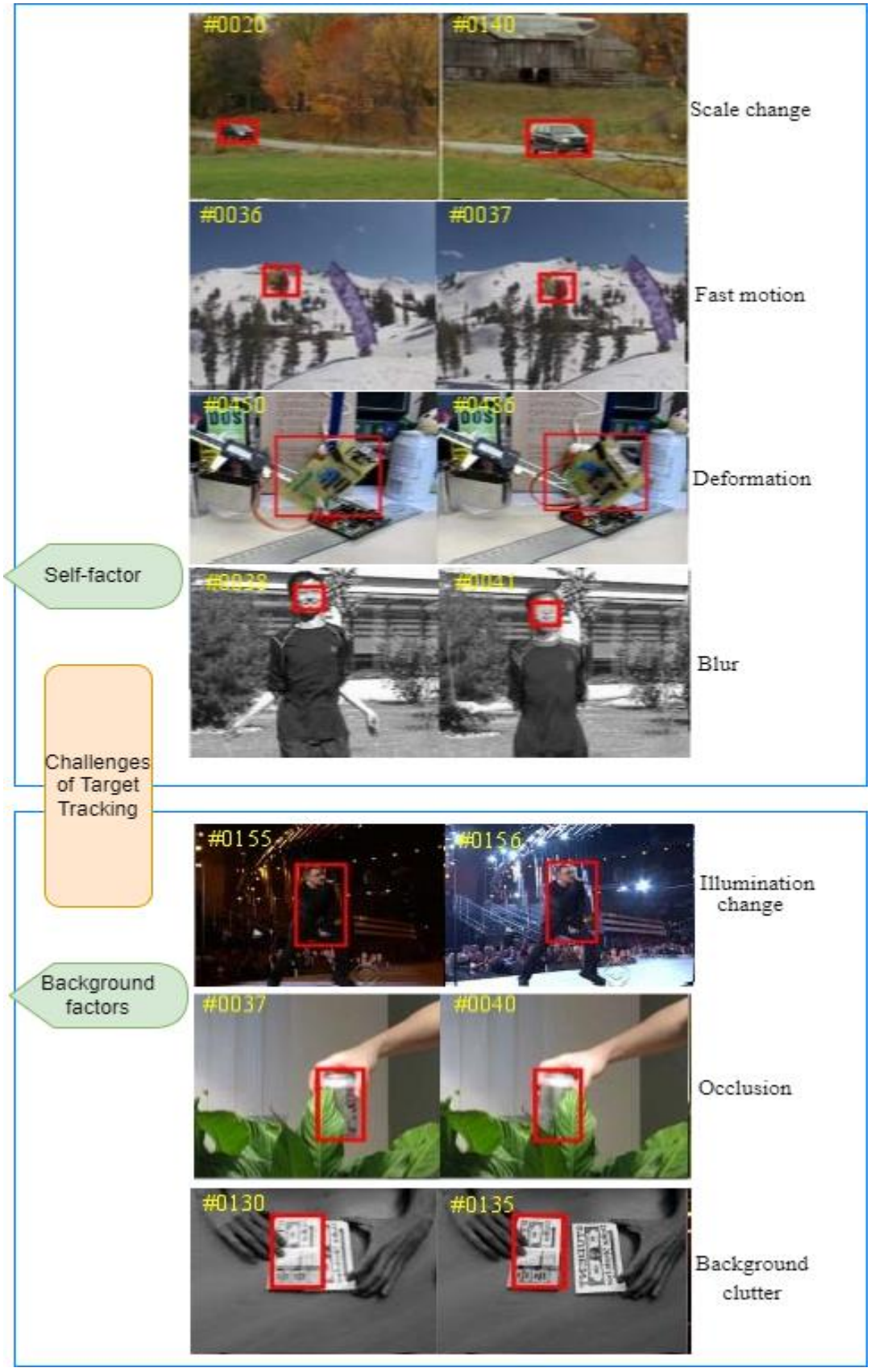

Figure 1. Challenging factors in target tracking.

Wang et al.[1] summarized the general framework of the target tracking system into five main parts, which are motion model, feature extraction, observation model, model update, and integration processing as shown in Figure 2. The motion model generates the target candidate region for the current frame; the feature extraction performs feature extraction on the candidate region, which is used to describe the properties of the target; the observation model determines whether the candidate region contains the target and uses it as the predicted target location; the model update is used to control the strategy of observation model update; and processing operation fuses the outputs of multiple sub-tracking algorithms to obtain the final output (multi-target tracking algorithm).

The literature[5-8] surveyed visual target tracking algorithms from different perspectives, but due to the rapid development of visual target tracking algorithms, especially based on the technical breakthroughs in deep learning tracking algorithms, there still needs to be more focused and comprehensive visual single-target tracking algorithms. Thus, this paper aims to provide a review of the research progress of visual single-target tracking methods based on basic deep learning theory, hoping to provide an organized and hierarchical reference of diverse single-target tracking algorithms and valuable ideas for future research work.

The work in this paper is organised as follows: Section 2 introduces traditional target tracking algorithms; Section 3 analyses the mainstream correlation filter-based video target tracking algorithms; Section 4 explores deep learning-based video target tracking algorithms; Section 5 covers the experiments; and Section 6 includes data analysis, results and future directions.

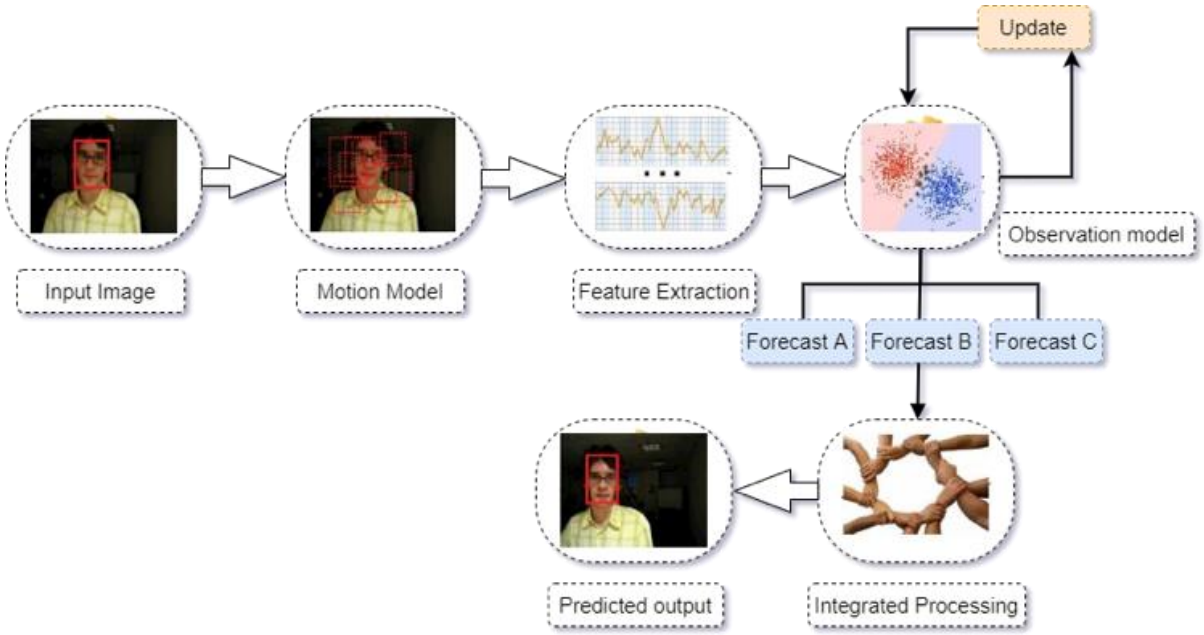

Figure 2. Target tracking process[1].

2 Research on Classical target Tracking Algorithm

Visual target tracking methods may be broadly categorized as either generative model-based or discriminative model-based target tracking algorithms. Generative model-based target tracking techniques are more common. Generative model-based target tracking algorithms, which utilise the results of historical frames to generate statistical models used to describe target characteristics, can effectively deal with target loss during

tracking. But generative model-based methods usually ignore the background information around the target, while easily losing the target in the face of background confusion. Before 2010, target tracking algorithms generally used generative model approaches, and the classical tracking algorithms were Meanshift, Particle Filter, Kalman Filter, and feature point-based optical flow algorithms, to name a few.

Unlike the generative model, the discriminative model mainly learns a decision boundary, which is used to distinguish the target region from the background region. After 2010, target tracking algorithms are mainly based on discriminative methods, and such discriminative classifiers are applied to target tracking algorithms; for example, support vector machines (SVMs), boosting algorithms, and decision trees, have achieved better results.

3 Research on Target Tracking Algorithm Based on Correlation Filtering

Since 2010, correlation filter (CF)-based tracking algorithms have gained popularity in academia and industry for their excellent performance and faster running speed, which have developed rapidly. Bolme et al.[9] proposed the minimum output sum of squared error (MOSSE) tracking algorithm, which was the first to introduce a correlation filter model in the field of target tracking to find the best position in subsequent frames by minimizing the mean-squared error In 2012, Henriques et al.[10] proposed the cyclic structure detection tracking algorithm with kernel (CSK), which uses a cyclic shift to densely sample the data and quickly train a classifier by fast Fourier transform (FFT). A number of related filtering algorithms have followed and built on them with a series of improvements in terms of feature representation, scale improvement, and resolution of boundary effects. Table 1 shows the technical comparison of various mainstream CF tracking algorithms.

3.1 Feature Improvement

Henriques et al.[10] extended the multi-channel function and kernel method based on CSK and proposed the Kernel Correlation Filter (KCF) tracking algorithm, while transforming the solution of correlation filter into a ridge regression problem. Danelljan et al.[11] mapped the original RGB 3-channel image to 11 channels in the Colour Name (CN) tracking algorithm and processed each channel individually before fusing the results. To solve the problem of too many channels affecting the running speed, principal component analysis (PCA) is used to reduce the dimensionality of two major channels from the 11 channels for the above processing. The efficient convolution operators for tracker with hand-crafted feature (ECO-HC)[12] using histogram of orientation gradient (HOG) and colour name (CN)[10] features were fused and good results were achieved.

Bertinetto et al.[13] proposed the sum of template and pixel-wise learners (STAPLE) tracking algorithm where HOG features and colour histogram are used to model the appearance of the target, that consist of some complementary features. By solving their response maps independently, a better tracking effect is obtained through a weighted fusion of the response maps.

Deep learning has achieved unprecedented results in the field of computer vision. In recent years, deep learning has also been introduced into the field of target tracking, where depth features are used to improve tracking performance under the tracking framework of correlation filtering.

3.2 Scale Improvement

Danelljan et al.[14] proposed the discriminative scale space tracker (DSST) algorithm, which views target tracking as two separate problems of target centre translation and scale change. The HOG feature is used to train the translation filter and the scale filter. The translation filter is used to obtain the target centre position, while the scale filter is used to calculate the confidence map. The scale corresponding to the response map that finds the

maximum value of the response is the best scale. In order to better cope with the scale variation, 33 scale filters were used, and this scaling method is also followed in the subsequent paper of Denlljan et al[14].

To better cope with scale changes, Li et al.[15] proposed the scale adaptive with multiple features tracker (SAMF) algorithm, which uses HOG features and CN features to extract features and seven scales for the target in the candidate region. This further detects both target translation changes and scale changes to determine the location and scale of the target quickly.

3.3 Handling Boundary Effects

Danelljan et al.[16] proposed the spatially regularized discriminative correlation filter (SRDCF) tracking algorithm to suppress the boundary effect by learning the correlation filter with larger spatial support in the detection phase. It maintains an extensive search range to better cope with the fast motion of the target.

The MOSSE-based correlation filters with limited boundaries (CFLM) tracking algorithm[16] and the background-aware correlation filters (BACF) tracking algorithm based on HOG features were proposed by Galoogahi et al.[17] Filters (BACF) tracking algorithm[17,18], is more effective in mitigating boundary effects by using larger size detection image blocks and smaller size filters to increase the proportion of real samples.

Unveiling the Power of Deep Tracking (UPDT) algorithm follows the Gaussian distribution used in ECO to extract positive samples, and also separates deep and shallow features. Experiments found that different features should be used with different variances. The influence of deep and shallow features in target tracking was systematically analysed and it was found that the deep model should be responsible for the robustness of the network while the shallow model was responsible for accurate localization. A novel feature fusion strategy is then proposed.

Danelljan et al.[19] proposed the ATOM tracking method, by designing a novel architecture consisting of specialized target estimation and classification components. An online trained classifier and an offline trained evaluation network were proposed to jointly solve the target tracking problem, which is very similar to detection, a two-stage tracking framework.

The tracking method of Probabilistic Regression for Visual Tracking (PrDiMP)[20] introduces meta-learning to incorporate the information of the first frame into the later frames, i.e., the information of the first frame is used to provide weights for the online update model of the later frames, where the online update model refers to the two Head parts of position prediction and bounding box prediction. Categorized as a regression problem, a conditional probability model is used here to predict the position of the next frame from the information of the previous frame.

Table 1. Comparison of the mainstream CF tracking algorithms.

F-Trackers	Features	Scale estimate	Offline training	Online learning
MOSSE[9]	Raw pixels	√	×	√
SAMF[15]	Raw pixels\HOG\CN	√	×	√
KCF[10]	Raw pixels\HOG	×	×	√
HCF[25]	HOG	×	×	√
DeepSRDCF[22]	HOG\CN	×	×	√
DSST[14]	HOG	√	×	√
STAPLE[13]	HOG\Colour histogram	√	×	√
ECO[12]	CNN\HOG\CN	√	×	√
UPDT[46]	CNN\HOG\CN	×	×	√
ATOM[19]	CNN	×	√	√
PrDiMP[20]	CNN	√	×	√

4 Research on Target Tracking Algorithm Based on Deep Learning (DL)

Deep learning-based target tracking algorithms can be divided into depth feature-based target tracking algorithms, Siamese network-based target tracking algorithms, recurrent neural network (RNN)-based target tracking algorithms, generative adversarial network (GAN)-based target tracking algorithms and other specific network-based target tracking algorithms. Table 2 shows the technical comparison of various mainstream DL tracking algorithms

4.1 Depth Feature-Based Target Tracking Algorithms

In depth feature-based target tracking algorithms, scholars have replaced the traditional features with depth features under the existing target tracking framework[21]. In 2015, Danelljan et al.[22] proposed the SRDCF framework using DeepSRDCF, an improved algorithm for feature extraction by VGGNet[23], which achieved better results, and also explored the effect of features of different layers of convolutional neural networks on target tracking accuracy. In 2016, Danelljan et al. also proposed the C-COT tracking algorithm, which uses VGGNet[23,24] to extract multi-resolution features in the continuous domain, interpolates multi-resolution features, and trains continuous correlation filters, which was used in the VOT2016 challenge, and resulted in an amazing performance. In 2017, Danelljan et al.[12] also proposed the ECO algorithm based on C-COT combining convolution features, HOG features and CN features by factorizing the ECO, which combines convolutional features, HOG features and CN features, reduces the dimensionality of features by factorization of convolution operations, and reduces the training samples in the learning model to improve the tracking speed and robustness.

Ma et al.[25] proposed the tracking algorithm of Hierarchical Convolutional Features for Visual Tracking (HCF) which uses three correlation filters. Since the upper layer provides semantic information and the bottom layer provides texture information, the correlation filters are used in the order from deep to shallow to determine the target location from coarse to fine.

4.2 Siamese Network-Based Target Tracking Algorithm

Scholars have suggested the use of Siamese network-based target tracking algorithms to overcome the poor speed caused by pre-trained networks as feature extractors. With quicker speed and greater tracking performance, Siamese networks have received much interest in target tracking.

Held et al.[26] suggested GOTURN in 2016. GOTURN introduced Siamese networks to target tracking and employed an offline feedforward network where a block of pictures from the current and previous frames is fed into a convolutional neural network for feature extraction and subsequently cascaded into a fully connected layer. The layer compares target and frame information to determine the target's location offset. Fully linked layer learns a complicated feature comparison function and outputs target motion.

Tao et al.[27] proposed Siamese Instance Search for Tracking (SINT) algorithm based on Siamese networks. SINT trains a matching function offline through a large amount of video data, which matches a given target in the initial frame with the next SINT trains a matching function offline by using a large amount of video data to match a given target in the initial frame with a candidate target in the next frame, and then returns the most similar target.

Bertinetto et al.[28] introduced Fully-Convolutional Siamese Networks for Object Tracking (SiamFC), which implements a fully convolutional Siamese network architecture and uses AlexNet as the backbone network to extract template and search picture features. The feature map of the template image is convolved with the feature map of the search image to create the response map. Figure 3 shows SiamFC's tracking architecture.

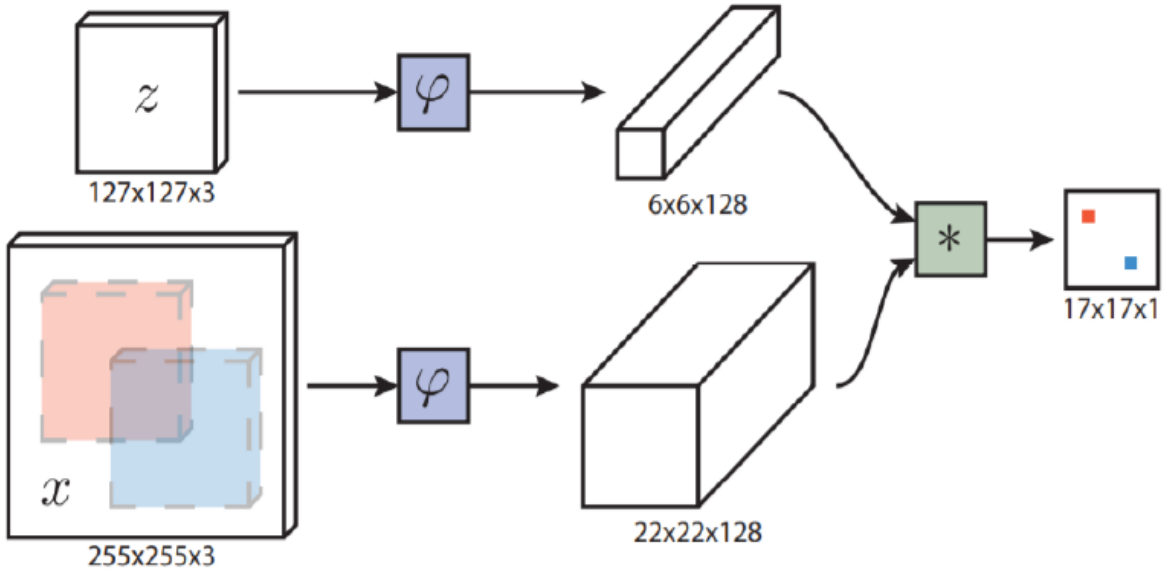

Figure 1. Tracking framework of SiamFC.

Valmadre et al.[29] improved on SiamFC to obtain the end-to-end representation learning for Correlation Filter based tracking (CFNet) algorithm. CFNet integrates correlation filtering into a network layer and adds to the template branch to update the template model, thus making the Siamese network more robust to appearance changes.

Li et al.[30] introduced Region Proposal Network (RPN)[31] to Siamese network target tracking and proposed SiamRPN tracking algorithm. SiamRPN is first trained end-to-end using large-scale images offline. In the tracking phase, the tracking task can be viewed as a single-sample detection task that directly regresses the target to be tracked without the need for scale estimation, greatly increasing the runtime speed.

SiamRPN++[32] presents a Depth-wise convolution design that saves arithmetic power without sacrificing accuracy. SiamRPN adds bounding box regression and short-term monitoring is restricted.

The Recurrently Optimizing Tracking Model (ROAM)[33] technique provides a tracking model with a resizable response generator and a bounding box modulator. Only one anchor size is utilized for each spatial location, and its convolution filter may adapt to shape changes through bilinear interpolation. A meta-learning-trained recurrent neural optimizer speeds up convergence of the updated tracking model.

The Siamese Fully Convolutional Classification and Regression for Visual Tracking (SiamCAR)[34] technique converts the network's regression output into a feature map using an anchor-free approach. Classification and centrality score maps are used to determine the optimum target centroid. The distance between the best target centroid and the four edges of the chosen box determines the tracking prediction box.

Siamese Box Adaptive Network for Visual Tracking (SIamBAN)[35] is built on Siamese network architecture and uses an anchor-free method, which gives the frame greater flexibility. Anchor-free removes predetermined anchors, which reduces model parameters and speeds it up. Null convolution improves perceptual field and tracking performance.

4.3 Recurrent Neural Network-Based Target Tracking Algorithm

Visual tracking is strongly tied to the spatial and temporal information of video frames, hence recurrent neural networks are progressively included in target tracking.

Structure-Aware Network for Visual Tracking (SANet)[36] is based on recurrent neural networks. The SANet employs RNNs to encode the structure of targets

throughout the learning process, which enhances target identification and interference source recognition. To supply richer information to the network, a layer-hopping connection method fuses CNNs and RNNs, and the algorithm's superior tracking effect is tested.

Yang et al.[37] proposed Learning Dynamic Memory Networks for Object Tracking (MemTrack). MemTrack is a dynamic memory network for visual tracking. The external storage unit is managed by a long-short term memory (LSTM) network with an attention mechanism to adjust to target appearance changes. Gated residual template learning generates the final matching template and prevents excessive model updating.

The SiamR-CNN[38] algorithm uses a hard case mining strategy to discriminate the interferers and designs a dynamic trajectory planning algorithm (TDPF) by which all object candidate frames in the previous frame are redetected and grouped into small trajectories over time, thus tracking all potential objects, including interferers, simultaneously. Then the best target is selected within the current time step using dynamic planning based on the complete history of all target and interfering object trajectories. Therefore, the algorithm is computationally intensive and cannot be tracked in real-time.

4.4 Generative Adversarial Network-Based Target Tracking Algorithms

Generative Adversarial Networks (GANs) have been extensively employed in various study domains to capture statistical distributions and generate training samples with little or labelled input.

Song et al.[39] applied GAN to target tracking and created an adversarial learning-based approach (VITAL). VITAL employs a generative network to randomly build masks and adaptively delete certain input attributes to boost positive samples. VITAL's network uses adversarial learning to identify masks that keep target object properties over time. VITAL presents a higher-order cost-sensitive loss to lessen the influence of clearly discernible negative samples while enabling network training.

Table 1. Comparison of the mainstream deep learning (DL) tracking algorithms.

Classification	Methods	Year	Filter	Offline training	Online learning	Features
Depth features	DeepSRDCF[22]	2016	CF	×	√	VGGNet
	C-COT[24]	2016	CF	×	√	VGGNet
	ECO[12]	2017	CF	×	√	HoG+CN+DL
	HCF[25]	2015	CF	×	√	VGGNet
SN	GOTURN[26]	2016	DL	√	×	VGGNet
	SINT[27]	2016	DL	√	×	AlexNet VGG16t
	SiamFC[28]	2015	DL	√	×	AlexNet
	CFNet[29]	2017	DL+CF	√	×	AlexNet
	SiamRPN[31]	2018	DL	√	×	AlexNet
	SiamRPN++[32]	2019	DL	√	×	AlexNet Resnet-50
	ROAM[33]	2020	DL	√	×	DAF
	SiamCAR[34]	2020	DL	√	×	Resnet-50
	Siamban[35]	2020	DL	√	×	Resnet-50
RNN	SANet[36]	2017	DL	√	×	R-CNN
	MemTrack[37]	2018	DL	√	×	R-CNN
	SiamR-CNN[31]	2019	DL	√	×	Fast R-CNN
GAN	VITAL[39]	2018	DL	√	×	GAN
Other	MDNet[40]	2016	DL	√	×	DAF
	TransT[41]	2021	DL	√	×	DAF

4.5 Target Tracking Algorithms Based on Other Specific Networks

Some researchers have created target tracking networks. Nam et al.[40] suggested Multi-Domain Convolutional Neural Network (MDNet) tracking technique. MDNet needs pre-training with several tracking movies to achieve a generic target representation. Each domain corresponds to a training sequence, and the shared layer learns the generic target representation during training. When a new sequence has to be updated, only the domain-specific layers of MDNet are updated online, allowing the network to adapt to the current tracking environment.

Chen et al.[41] introduced a novel Transformer tracking system, including feature extraction, class fusion, and head prediction modules. Transformer class fusion mixes template and searches region characteristics without correlation. Feature fusion networks based on self-context enhancement and cross-feature enhancement are created, focusing on important information, including edges and comparable targets, as well as building correlations between distant data that improves classification and regression outcomes.

5 Experimental

This section gives experimental data on the performance of the two types of target tracking algorithms discussed in Sections 3 and 4 on the OTB100, and VOT2018 benchmark datasets. Table 3 gives the details of some common single-target tracking benchmark datasets.

5.1 Evaluation Methods for Single Target Tracking

To promote the development of the target tracking field, scholars have summarized and generalized the evaluation criteria of target tracking algorithms, i.e., the performance of different tracking algorithms is evaluated by qualitative and quantitative evaluations. For qualitative analysis, three evaluation criteria are commonly used: traditional evaluation methods, Visual Object Tracking (VOT) evaluation methods[42,43] and Online Object Tracking Benchmark (OTB) evaluation methods[3,4].

5.1.1 Traditional Evaluation Methods

The traditional evaluation methods include two metrics, central location error (CLE) and overlap ratio (OR)[44]. The smaller the CLE value, the higher the accuracy of the algorithm. The larger the OR value, the better the tracking performance of the algorithm. Generally, using the average overlap rate in the tracking algorithm can reflect the tracking accuracy more accurately.

5.1.2 OTB Evaluation Methods

On the basis of a description of prior work, Wu et al.[3,4] presented the target tracking benchmark OTB for assessing the performance of single-target tracking algorithms. The OTB assessment database originally had 50 video sequences, and the OTB not only offers evaluation metrics for testing target tracking systems, but also includes some well-labelled, challenging video sequences. The OTB benchmark also offers an assessment toolkit with MATLAB and Python versions, and the function interface is straightforward and easy to use. Therefore, it is frequently used. The OTB analyses the performance of the tracking algorithm using the precision rate (PR) based on the centre position error and the accuracy rate based on the target tracking method, and the success rate (SR) is determined by the overlap rate.

The success rate chart of the algorithm can be developed based on the success rate of the target tracking algorithm under different thresholds The area under the curve (AUC) of the success rate chart is used to rank different tracking algorithms and compare the advantages and disadvantages of the algorithms, based on the accuracy rate. metric based on the centre position error and the success rate metric based on the overlap rate. OTB proposes three metrics: one pass evaluation (OPE), temporal robustness evaluation (TRE), and spatial robustness evaluation (SRE). These three values represent the PR and SR of different tests.

The larger the value, the better the tracking accuracy and tracking performance.

The performance of the target tracking algorithm can be easily evaluated by the OTB evaluation method, using metrics such as accuracy and success rate to assist in analysing the performance of the algorithm, as well as to evaluate and compare different algorithms.

5.1.3 VOT Evaluation Methods

Since 2013, VOT has been an annual target tracking competition[43,44,46] that typically acts as a workshop for IEEE International Conference on Computer Vision (ICCV) and European Conference on Computer Vision (ECCV) conferences. The number of test video sequences on VOT has climbed from 16 in 2013 to 60 currently, while the complexity of the video sequences has constantly increased. Since VOT provides resources such as evaluation criteria required to assess the performance of tracking algorithms, a large number of manually labelled test videos, open source evaluation toolkits, and test results of many tracking algorithms on VOT, the VOT evaluation method has been widely adopted in the field of target tracking.

Starting from VOT2016, three key metrics to evaluate the performance of target tracking algorithms are used in VOT: accuracy (A), robustness (R), and expected average overlap (EAO). The larger the accuracy value, the higher the tracking accuracy. The smaller the robustness value, the better the tracking performance. The larger the EAO value, the higher the target tracking accuracy.

In addition to the above experimental datasets, a number of others have emerged in recent years, such as UAV123, LaSOT[45], as shown in Table 3.

Table 2. Video Count (VC), Minimum Frame Rate (Min-FR), Maximum Frame Fate (Max_FR), and Total Frames (TF) of evaluation datasets for major single-target tracking.

DATASET	VC	Min-FR	Max-FR	TF	LINK
OTB50	50	71	3,872	29,491	http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html
OTB100	100	71	3,872	59,040	http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html
VOT2018	60	41	1,500	21,356	http://www.votchallenge.net/vot2018/
UAV123	123	109	3,085	113,000	https://cemse.kaust.edu.sa/ivul/uav123
LaSOT	1,400	1,000	11,397	3,520,000	https://cis.temple.edu/lasot/download.html

5.2 Experimental Data of Target Tracking Algorithm

To get an accurate understanding of the performance of the classical single-target tracking algorithm, we tested on a high-performance computer with an Inteli7-12700H CPU and paired with a GeForce RTX3070 and 32G RAM, with data sets based on OTB100 and VOT2018, ranging from the classical single-target tracking algorithm. In OTB100, two metrics, PR and AUC, were used to measure the performance of the algorithms, PR is the accuracy rate based on the center position error, and AUC is the area under the curve through the success rate plot to rank and compare the different tracking algorithms for the algorithms' merits. Both values were taken as the average of 11 attributes in the OTB100, and higher values represent better corresponding performance of the algorithms. In VOT2018, three metrics, A, R, and EAQ, were used to compare the performance of the algorithms. A stand for accuracy, the tracking frame predicted by the target tracking algorithm in the test video, and the overlap between the predicted target bounding box and the manually marked target bounding box was calculated. In contrast, the performance of the algorithm was measured by the degree of overlap of the bounding box. The higher the overlap rate, the better the accuracy of the target tracking algorithm. R stands for robustness, where the target tracking algorithm may not succeed in a single run after the test video. It may need several re-initializations to succeed, which depends on the number of times the algorithm is

reinitialized to characterize the robustness of the target tracking algorithm. The lower the number of re-initializations, the better the robustness of the algorithm. The larger the EAO value, the higher the accuracy of the target tracking algorithm. All algorithm codes are available at the official download source published by the algorithm founders, and the speeds of the algorithms are obtained from the officially published data.

6 Results and Discussion

6.1 Experimental Data Analysis

From Table 4**,** Figure 4 and Figure 5, it can be concluded that the discriminative tracking approach converts the tracking problem into a detection problem when analysed from the perspective of features, so good features are the key factor for such tracking. From the success rate and accuracy results given in Table 5, it can be seen that HOG and CN features reflect excellent performance in the field of visual tracking, and many methods proposed afterwards combine depth features in different ways to construct tracking frameworks that reflect good performance. The biggest advantage of the correlation filtering-based tracking algorithm is reflected in the speed.

As can be derived from Table 5, Figure 6 and Figure 7, the results show that the MDNet algorithm based on video data trained offline and with online model updates, alongside the improved MDNet-based algorithm VITAL achieved good results in terms of tracking accuracy, but was not satisfactory in terms of speed and did not meet the real-time criteria. The algorithms SiamFC, Dsiam and SINT based on the Siamese network framework also achieved relatively good rankings. C-COT uses VGG-Net to extract depth features, using the original colour image and the output of two convolutional layers as features, which have significantly improved accuracy compared with similar algorithms. Still, the various features seriously reduce the computational efficiency and make it challenging to meet the real-time requirements. ECO reduces the feature dimensions of HOG, CN and CNN by factorization operation on the basis of C-COT, where HOG is compressed to 10, CN is compressed to 3. The 1st and 5th convolutional layers of CNN are compressed to 16 and 64, respectively, reducing the training parameters and thus, effectively reducing the computational complexity. The tracking performance is very high in the experimental data for each dataset.

Table 3. Experiment mainstream CF tracking algorithms in OTB100, VOT2018.

CF-Trackers	OTB100		Speed (fps)		VOT2018		
	PR	AUC	CPU	GPU	EAO	A	R
MOSSE[19]	0.421	0.308	669.0	-	0.128	0.499	0.962
SAMF[14]	0.723	0.547	7.0	-	0.091	0.464	1.292
KCF[9]	0.696	0.465	172.0	-	0.129	0.451	0.768
HCF[20]	0.837	0.562	-	10.4	-	-	-
DeepSRDCF[21]	0.849	0.651	-	0.2	0.151	0.485	0.703
DSST[13]	0.725	0.552	54.3	-	-	-	-
STAPLE[12]	0.782	0.581	80.0	-	0.169	0.469	0.599
ECO[11]	0.929	0.706	60.0	8.0	0.301	0.467	0.267
UPDT[22]	0.928	0.701	-	-	0.367	0.527	0.175
ATOM[23]	0.881	0.657	-	30.0	0.399	0.587	0.201
PrDiMP[18]	0.911	0.701	-	30.0	0.438	0.612	0.157

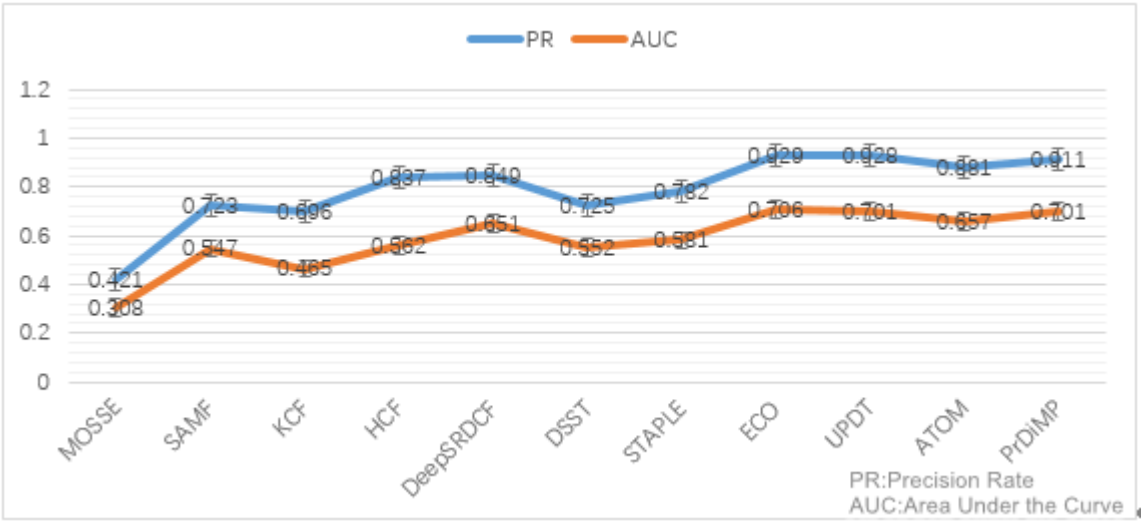

Figure 2. Comparison of performance of mainstream CF trackers in OTB100.

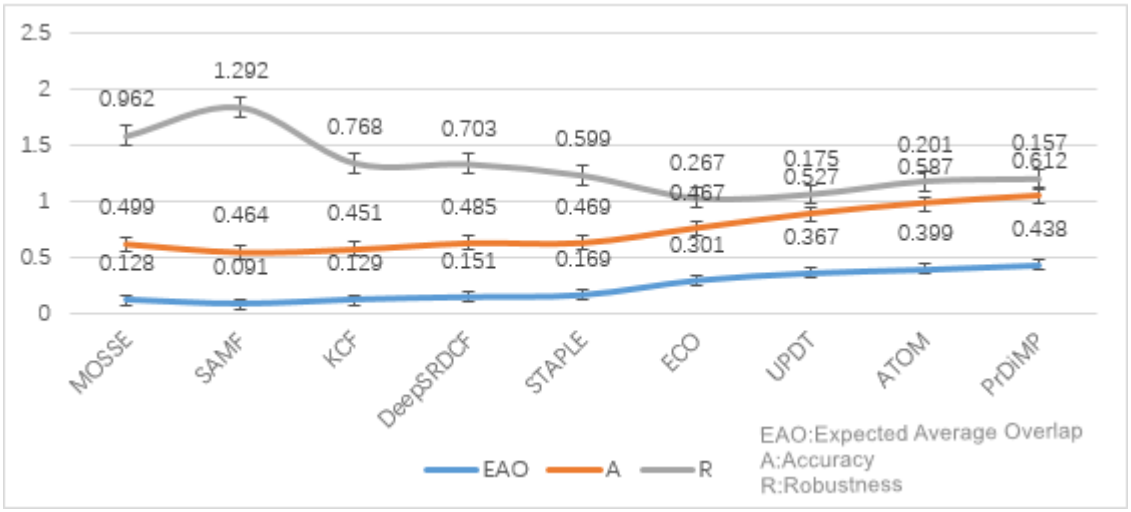

Figure 3. Comparison of performance of mainstream CF trackers in VOT2018.

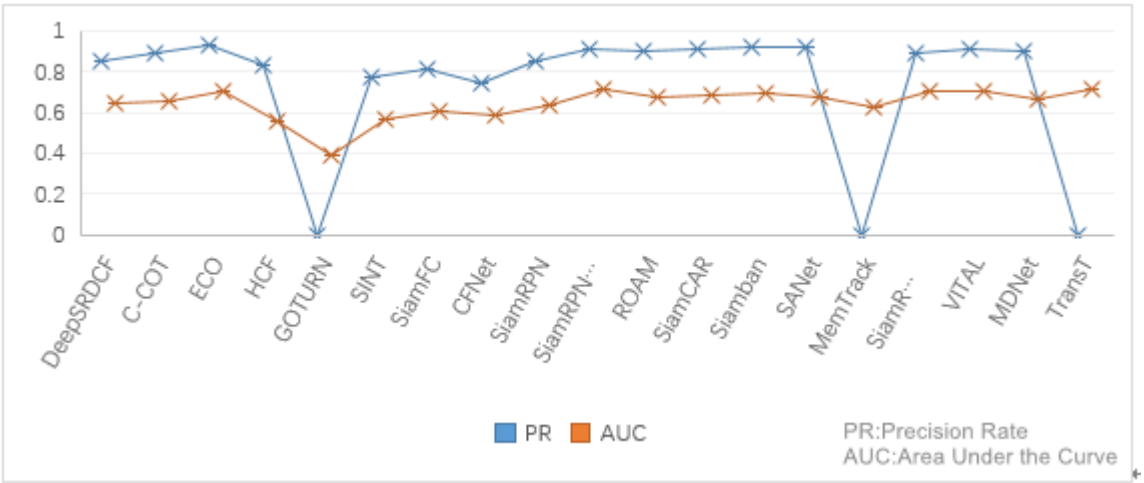

Figure 4. Comparison of performance of mainstream DL trackers in OTB100.

Table 4. Experiment mainstream DL tracking algorithms in OTB100, VOT2018.

Classification	Methods	VOT2018			OTB100		
		EAO	A	R	PR	AUC	FPS
Depth Features	DeepSRDCF[21]	0.276	0.528	0.326	0.849	0.651	0.3
	C-COT[26]	0.331	0.539	0.238	0.891	0.659	0.3
	ECO[11]	0.374	0.553	0.200	0.929	0.706	6.0
	HCF[20]	0.395	0.561	0.199	0.837	0.562	0.8
SN	GOTURN[27]	0.240	0.390	0.100	-	0.390	100.0
	SINT[28]	0.201	0.506	0.231	0.778	0.571	68.0
	SiamFC[29]	0.236	0.534	0.541	0.815	0.612	58.0
	CFNet[30]	0.186	0.501	0.585	0.749	0.591	75.0
	SiamRPN[32]	0.344	0.560	0.260	0.849	0.637	126.0
	SiamRPN++[33]	0.414	0.600	0.234	0.912	0.715	35.0
	ROAM[34]	0.380	0.543	0.195	0.902	0.680	20.0
	SiamCAR[35]	-	-	-	0.907	0.689	52.0
	Siamban[36]	0.452	0.597	0.178	0.917	0.696	40.0
	SANet[37]	0.389	0.610	0.690	0.926	0.677	1.0
RNN	MemTrack[38]	0.273	0.530	0.440	-	0.628	50.0
	SiamR-CNN[32]	0.408	0.609	0.220	0.890	0.701	4.7
GAN	VITAL[40]	0.323	0.630	0.170	0.911	0.710	1.5
Other	MDNet[41]	0.211	0.600	0.160	0.905	0.670	1.0
	TransT[42]	-	-	-	-	0.711	50.0

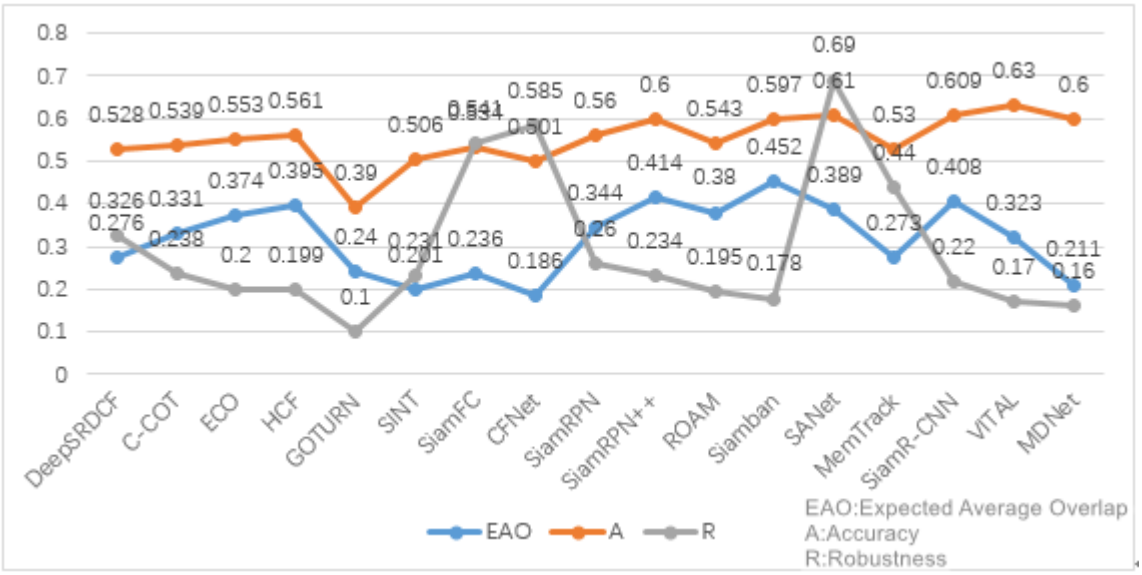

Figure 5. Comparison of performance of mainstream DL trackers in VOT2018.

6.2 Results

The results show that ECO is the best performing CF algorithm in terms of tracking accuracy, tracking speed, and other aspects of performance.

Among the DL algorithms, the Siamese network structure of the algorithm alone is not particularly outstanding in all aspects, but it is the best in terms of stability. In particular, combined with the use of the lightweight network model SANet[35], the comprehensive performance in various aspects such as tracking accuracy and tracking speed is the best.

6.3 The Future Direction of Development

The future works could be in two main directions. First, how to balance the relationship between tracking performance and real-time. Mainly in the balance between tracking accuracy and tracking speed. If the accuracy of the algorithm is

good, but cannot be used for real-time, it cannot be converted into products.

Second, visual saliency, attention mechanism and the integration of various modules of target tracking, weakening the background to highlight the foreground, guiding the tracker to focus on useful information, and realizing the combination of correlation filtering and twin networks will all be the space for researchers to explore.

7 Conclusion

This paper focuses closely on the visual target tracking framework. Firstly, the traditional visual target tracking algorithm was analysed. Then, the mainstream video target tracking algorithms based on correlation filtering were analysed from three aspects: feature improvement, scale improvement, and dealing with boundary effects. Then, the video target tracking algorithms based on deep learning were discussed in detail, and the target tracking algorithms based on deep learning were divided into five major categories, and each type of algorithm was analysed in terms of research motivation, algorithmic ideas, research framework, advantages and disadvantages. Finally, the tracking algorithms analysed above have experimented on OTB100 and VOT2018 benchmark datasets, and the experimental data obtained were compared to draw the authors' conclusions on visual single-target tracking algorithms and point out the future development trend in the field of video target tracking.

Conflict of Interest

No potential conflict of interest was reported by the authors.

Acknowledgment

ZhongMing Liao acknowledges the support of the InnoSTRE 2022 conference organising committee.

Funding

No funding sources.

Author Contribution

Conceptualization: Liao, Z.
Data curation: Liao, Z.
Methodology: Liao, Z.
Formal analysis: Liao, Z.
Visualisation: Liao, Z.
Software: Liao, Z.
Writing (original draft): Liao, Z.
Writing (review and editing): Ismail, A.
Validation: Ismail, A., & Liao, Z.
Supervision: Ismail, A.
Funding acquisition: Not applicable
Project administration: Ismail, A.

References

1. Wang, N., Shi, J., Yeung, D. Y., & Jia, J. (2015, December). Understanding and diagnosing visual tracking systems. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV2015)* (pp. 3101-3109). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.355
2. Qi, Y., Zhang, S., Qin, L., Huang, Q., Yao, H., Lim, J., & Yang, M. H. (2018). Hedging deep features for visual tracking. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *41*(5), 1116-1130. https://doi.org/10.1109/TPAMI.2018.2828817
3. 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). https://doi.org/10.1109/CVPR.2013.312
4. Wu, Y., Lim, J., & Yang, M. H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1834-1848. https://doi.org/10.1109/TPAMI.2014.2388226
5. Hechun, W., & Xiaohong, Z. (2019, August). Survey of deep learning based object detection. In *Proceedings of the 2nd International Conference on Big Data Technologies* (pp. 149-153). https://doi.org/10.1145/3358528.3358574
6. Li, P., Wang, D., Wang, L., & Lu, H. (2018). Deep visual tracking: Review and experimental comparison. *Pattern Recognition, 76,* 323-338. https://doi.org/10.1016/j.patcog.2017.11.007
7. Jiao, L., Wang, D., Bai, Y., Chen, P., & Liu, F. (2021). Deep learning in visual tracking: A review. In *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2021.3136907.
8. Marvasti-Zadeh, S. M., Cheng, L., Ghanei-Yakhdan, H., & Kasaei, S. (2022). Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems, 23*(5), 3943-3968. https://doi.org/10.1109/TITS.2020.3046478

9. 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 (CVPR2010),* (pp. 2544-2550). IEEE. https://doi.org/10.1109/CVPR.2010.5539960
10. Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2014). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI2015), 37*(3), 583-596. https://doi.org/10.1109/TPAMI.2014.2345390
11. Danelljan, M., Shahbaz Khan, F., Felsberg, M., & Van de Weijer, J. (2014). Adaptive color attributes for real-time visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CPRV2014)* (pp. 1090-1097). https://doi.org/10.1109/CVPR. 2014.143
12. Danelljan, M., Bhat, G., Shahbaz Khan, F., & Felsberg, M. (2017). ECO: Efficient convolution operators for tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CPRV2017)* (pp. 6638-6646). https://doi.org/10.1109/CVPR.2017.733
13. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., & Torr, P. H. (2016). STAPLE: Complementary learners for real-time tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016)* (pp. 1401-1409). https://doi.org/10.1109/CVPR.2016.156
14. Danelljan, M., Häger, G., Khan, F. S., & Felsberg, M. (2016). Discriminative scale space tracking. IEEE *Transactions On Pattern Analysis and Machine Intelligence (PAMI2016)*, 39(8), 1561-1575. https://doi.org/10.1109/TPAMI.2016.2609928
15. Li, Y., & Zhu, J. (2015, September). A scale adaptive kernel correlation filter tracker with feature integration. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II 13* (pp. 254-265). Springer International Publishing. https://doi.org/10.1007/978-3-319-16181-5_18
16. 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 (ICCV2015)* (pp. 4310-4318). doi: 10.1109/ICCV.2015.490. https://doi.org/10.1109/ICCV.2015.490
17. Galoogahi, K.H., Sim, T., & Lucey, S. (2015, June). Correlation filters with limited boundaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4630-4638). IEEE Computer Society. https://doi.org/10.48550/arXiv.1403.7876
18. Galoogahi, H. K., Fagg, A., & Lucey, S. (2017, October). Learning background-aware correlation filters for visual tracking. In *2017 IEEE International Conference on Computer Vision (ICCV2017)* (pp. 1144-1152). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.129
19. Danelljan, M., Bhat, G., Khan, F. S., & Felsberg, M. (2019, June). ATOM: Accurate tracking by overlap maximization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2019)* (pp. 4655-4664). IEEE. https://doi.org/10.1109/CVPR.2019.00479
20. Danelljan, M., Van Gool, L., & Timofte, R. (2020, June). Probabilistic regression for visual tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2020),* (pp. 7181-7190). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00721
21. Ojala, T., Pietikainen, M., & Harwood, D. (1994, October). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition* (Vol. 1, pp. 582-585). IEEE. https://doi.org/10.1109/ICPR.1994.576366
22. Danelljan, M., Häger, G., Khan, F. S., & Felsberg, M. (2015, December). Convolutional features for correlation filter based visual tracking. In *2015 IEEE International Conference on Computer Vision Workshop (ICCV-W2015)* (pp. 621-629). IEEE. https://doi.org/10.1109/ICCVW.2015.84
23. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *The 3rd International Conference on Learning Representations (ICLR2015)*. *https://arxiv.org/abs/1409.1556*
24. Danelljan, M., Robinson, A., Shahbaz Khan, F., & Felsberg, M. (2016). Beyond correlation filters: Learning continuous convolution operators for visual tracking. In B. Leibe, J. Matas, N. Sebe & M. Welling (Eds.), *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science* (Vol. 9909, pp. 472-488). Springer. https://doi.org/10.1007/978-3-319-46454-1_29
25. Ma, C., Huang, J. B., Yang, X., & Yang, M. H. (2015, February). Hierarchical convolutional features for visual tracking. In *15th IEEE International Conference on Computer Vision (ICCV 2015)* (pp. 3074-3082). Institute of Electrical and Electronics Engineers Inc. https://doi.org/ 10.1109/ICCV.2015.352
26. Held, D., Thrun, S., & Savarese, S. (2016, October). Learning to track at 100 fps with deep regression networks. In *Computer Vision–ECCV 2016: 14th European Conference Proceedings,* Amsterdam, The Netherlands (Part I 14, pp. 749-765). Springer. https://doi.org/10.48550/arXiv.1604.01802
27. Tao, R., Gavves, E., & Smeulders, A. W. (2016, June). Siamese instance search for tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016)* (pp. 1420-1429). IEEE Computer Society. https://doi.org/10.1109/CVPR.2016.158
28. Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016, October). Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops Proceedings,* Amsterdam, The Netherlands, (Part II 14, pp. 850-865). Springer. https://doi.org/10.48550/arXiv.1606.09549

29. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., & Torr, P. H. (2017, July). End-to-end representation learning for correlation filter based tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5000-5008). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.531
30. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(06), 1137-1149. https://doi.ieeecomputersociety.org/10.1109/TPAMI.2016.2577031*
31. Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018, June). High performance visual tracking with Siamese region proposal network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2018)* (pp. 8971-8980). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00935
32. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., & Yan, J. (2019, June). Siamrpn++: Evolution of siamese visual tracking with very deep networks In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4277-4286). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00441
33. Yang, T., Xu, P., Hu, R., Chai, H., & Chan, A. B. (2020, June). ROAM: Recurrently optimizing tracking model. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6717-6726). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00675
34. Guo, D., Wang, J., Cui, Y., Wang, Z., & Chen, S. (2020, June). SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6268-6276). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00630
35. Chen, Z., Zhong, B., Li, G., Zhang, S., & Ji, R. (2020, June). Siamese Box Adaptive Network for visual tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6667-6676). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00670
36. Fan, H., & Ling, H. (2017, July). SANet: Structure-Aware Network for visual tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 2217-2224). IEEE. https://doi.org/10.1109/CVPRW.2017.275
37. Yang, T., & Chan, A. B. (2018, September). Learning dynamic memory networks for object tracking. In *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IX* (pp. 153-169). https://doi.org/10.1007/978-3-030-01240-3_10
38. Voigtlaender, P., Luiten, J., Torr, P. H., & Leibe, B. (2020, June). Siam R-CNN: Visual tracking by re-detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2020)* (pp. 6577-6587). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00661
39. Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Lau, R.W.H., & Yang, M. H. (2018, June). VITAL: VIsual tracking via adversarial learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2018)* (pp. 8990-8999). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00937
40. Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4293-4302). https://doi.org/10.48550/arXiv.1510.07945
41. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., & Lu, H. (2021, June). Transformer tracking. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8122-8131). IEEE Computer Society. https://doi.org/ 10.1109/CVPR46437.2021.00803
42. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin, L., Vojíř, T., Häger, G., Lukežič, A., Fernández, G., Gupta, A., Petrosino, A., Memarmoghadam, A., Garcia-Martin, A., Montero, A. S., Vedaldi, A., Robinson, A., Ma, A. J., Varfolomieiev, A., …, & Chi, Z. (2016). The visual object tracking VOT2016 challenge results. In G. Hua & H. Jégou (Eds.), *Computer vision – ECCV 2016 workshops. ECCV 2016. Lecture Notes in Computer Science* (Vol. 9914). Springer. https://doi.org/10.1007/978-3-319-48881-3_54
43. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Zajc, L. Č., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A., Fernandex, G., Garcia-Martin, A., Iglesias-Arias, A., Alatan, A. A., Gonzalez-Garcia, A., Petrosino, A., Memarmoghadam, A., Vedaldi, A., Muhic, A., ..., & He, Z. (2019). The sixth visual object tracking VOT2018 challenge results. In L. Leal-Taixé & S. Roth (Eds.), *Computer vision - ECCV 2018 Workshops. European Conference on Computer Vision (ECCV) 2018. Lecture Notes in Computer Science* (Vol. 11129, pp. 3-53). Springer. https://doi.org/10.1007/978-3-030-11009-3_1
44. Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, *111*, 98-136. https://doi.org/10.1007/s11263-014-0733-5
45. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., & Ling, H. (2019, June). LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2019)* (pp. 5369-5378). IEEE. https://doi.org/10.1109/CVPR.2019.00552

46. Bhat, G, Johnander, J, Danelljan, M., Khan, F. S., & Felsberg, M. (2018). Unveiling the power of deep tracking. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, (Vol. 11206, pp. 493-509). Springer. https://doi.org/10.1007/978-3-030-01216-8_30