Automation, Control and Intelligent Systems
2024, Vol. 12, No. 2, pp. 35-47
https://doi.org/10.11648/j.acis.20241202.12

Science**PG**
Science Publishing Group

Research Article

UAV Visual Tracking with Enhanced Feature Information

Shuduo Zhao[1, *], Yunsheng Chen[2], Shuaidong Yang[1]

[1]School of Electric and Information, Southwest Petroleum University, Chengdu, China
[2]School of Automation, Chongqing University, Chongqing, China

Abstract

Unmanned aerial vehicles (UAVs) visual tracking is an important research direction. The tracking object is lost due to the problems of target occlusion, illumination variation, flight vibration and so on. Therefore, based on a Siamese network, this study proposes a UAVs visual tracker named SiamDFT++ to enhance the correlation of depth features. First, the network width of the three-layer convolution after the full convolution neural network is doubled, and the appearance information of the target is fully utilized to complete the feature extraction of the template frame and the detection frame. Then, the attention information fusion module and feature deep convolution module are proposed in the template branch and the detection branch, respectively. The feature correlation calculation methods of the two depths can effectively suppress the background information, enhance the correlation between pixel pairs, and efficiently complete the tasks of classification and regression. Furthermore, this study makes full use of shallow features to enhance the extraction of object features. Finally, this study uses the methods of deep cross-correlation operation and complete intersection over union to complete the matching and location tasks. The experimental results show that the tracker has strong robustness in UAVs short-term tracking scenes and long-term tracking scenes.

Keywords

Visual Tracking, Unmanned Aerial Vehicles, Convolutional Feature, Real-time Remote Sensing

1. Introduction

Owing to the strong flexibility and high safety performance of UAVs, the construction of UAVs air platforms has been continuously improved in recent years, and has been widely used in autopilot, human-computer interaction, target following and so on. UAVs visual tracking principally focuses on the target position in the first video frame, and then locates and predicts the target in the subsequent video frames. At present, the mainstream UAVs visual tracking algorithms are principally divided into three categories: 1) Classical tracking algorithms, such as optical flow and Kalman filter. 2) Correlation filter tracking algorithm, such as kernel correlation filters [2], staple [3]. 3) Deep learning tracking algorithms, such as fully convolutional siamese networks for object tracking [4] (SiamFC) and learning to track at 100 FPS with deep region networks [1].

UAVs visual tracking is a challenging task. Visual tracking based on an air platform will encounter many challenges, such as aspect ratio change, partial/full occlusion, low resolution, similar object and camera jitter. Before the development of deep learning, UAVs visual trackers mainly rely on manual features to extract the features of targets, such as HOG, CN and GRAY. Because of its high efficiency in the frequency

[*]Corresponding author: shuduozhao@foxmail.com (Shuduo Zhao)

Received: 5 July 2024; **Accepted:** 26 July 2024; **Published:** 15 August 2024

Copyright: © The Author(s), 2024. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

domain, the correlation filtering framework for visual tracking uses the cyclic matrix to generate negative sample information, so that the tracker can learn the context information and locate the target accurately. Although these trackers [2, 5-7] have greatly improved the tracking performance compared with traditional algorithms, due to the complexity of UAVs tracking scenes, these trackers still have poor robustness.

The UAVs visual tracker based on deep learning primarily uses the backbone network for feature extraction, including AlexNet, VGGNet, ResNet, etc., which effectively improves the tracking performance. SiamFC extracts features and completes similarity matching through a full convolution neural network to obtain the feature map of the target location, which effectively improves the robustness of the tracker. Object tracking algorithms based on deep learning have gradually become mainstream, including high performance visual tracking with siamese region proposal network [8] (SiamRPN), unsupervised deep tracking [9] (UDT), target aware deep tracking [10] (TADT) and so on. In the process of UAVs visual tracking, feature extraction through lightweight convolutional neural network (such as AlexNet) still contains considerable background information, which seriously affects the tracking performance of the tracker. The use of deeper networks will lead to the destruction of translation invariance in the convolution process [11], and the tracking speed cannot achieve real-time performance. The purpose of using expanded convolution [12, 13] is to expand the receptive field and avoid the resolution of degradation caused by the pooling layer, but this method still suffers from performance instability when dealing with small targets. The UAVs tracker based on deep learning [14-17] still has some limitations and cannot achieve a good balance between accuracy and speed. In addition, although many trackers can utilize deep neural networks to extract the information of target features, they seriously affect the real-time performance of the tracking process. The use of a lightweight network can effectively improve the tracking speed, but the lack of access to the information of the deep characteristics of the target leads to poor context relevance of the model. Therefore, many trackers do not fully consider the impact of the model itself on the tracking performance, resulting in low tracking accuracy. However, the tracker designed not only uses a lightweight network to complete the model training, but also enhances the relevance of the model to the pixel pair of the template frame and detection frame, and improves the ability of target positioning.

Although SiamRPN has achieved a remarkable tracking effect by introducing regression calculations, there are many instabilities in flight tracking in UAVs scenes, such as occlusion, background clutter, illumination variation and so on. Therefore, this study proposes the UAVs visual tracker SiamDFT++ (deep feature enhancement tracking of the Siamese network). AlexNet is used to complete the target feature extraction. The difference here is that the number of channels is doubled in the latter three-layer convolution to reinforce the appearance features related to the target. This study introduces CycleMLP [18] to focus on shallow object features with rich information. Meanwhile, this study uses the methods of deep cross correlation [11] operation and complete intersection over union [19] (CIOU) to complete the matching and location tasks. The main contributions of this work are as follows:

Because the target information is fixed in the process of UAVs tracking, an attention information fusion module (AIFM) is designed for the extraction of target features in the template frame. It is used to adaptively learn the target appearance features of the template frame, to improve the weight proportion of the target spatial position and channel information in the network. Due to the continuous change in detection frame information, it is worth considering how to dynamically strengthen the extraction of target feature information in the changing detection frame after extracting features from the backbone network. Therefore, a feature deep convolution module (FDCM) is proposed in this paper. It is used to learn the changing feature information, enhance the search and extraction ability of target information, and complete the final classification and regression task.

This study evaluates the SiamDFT++ on three authoritative aerial benchmarks. Through quantitative and qualitative analysis, the performance of the tracker is better than many state-of-the-art (SOTA) trackers.

The actual test on a typical air platform proves the superior efficiency and effectiveness of SiamDFT++ in real-world scenarios.

2. Related Work

Over the past few years, many target trackers have demonstrated good performance on any aircraft. Although learning the continuous convolution operator [15] (CCOT) of visual tracking uses vggnet to extract features and learns the discriminant convolution operator in continuous space, it basically completes feature extraction through image convolution. ECO [16] utilizes channel compression and model updating strategies to improve the robustness of the tracker. Although these trackers [10, 20-23] have achieved some improvements by using multi-feature fusion and subsequent fusion strategies to perform feature extraction, the ability of using global feature information for trackers still needs to be enhanced. Their robustness in executing UAVs visual tracking tasks is poor, and the ability of applying global information for these trackers still needs to be enhanced.

Siamese-based methods [4, 8, 10, 17] have obtained excellent tracking performance. SiamFC achieves faster tracking speed and higher accuracy. SiamRPN [8] adds a regression branch with fine-tuned bounding boxes to improve the tracking performance. SiamCorners [17] used the corner pool module to predict target corner points, and then conducted multi-level feature fusion to further predict multiple corner points after mutual attention, achieving good tracking performance. However, these trackers [4, 8, 10, 18, 24, 25] use only deep networks (e.g., ResNet and VGGNet) to accomplish feature extraction before calculating the correlation and do not sufficiently consider whether the features extracted from the current image can adequately characterize the extracted target feature information. At the same time, the use of deeper network seriously affects the real-time tracking for UAVs. Therefore, this study designed AIFM and FDCM to alleviate the above problems.

3. UAV Visual Tracking Algorithm

3.1. Overall Framework of the Algorithm

In recent years, UAVs visual tracking has been broadly applied in both military and civilian applications. Aerial vehicles are deployed with visual sensors to perform visual tracking tasks. An increasing number of researchers are working on small UAVs and applying them to diverse fields. However, due to the complex background of UAVs tracking process, there are multiple challenges. The development of an efficient and robust visual tracking method is fundamental for the future of UAVs remote sensing applications.

The tracker designed can conduct end-to-end learning. The baseline tracker [8] also does not make full use of the target's shallow feature information. The baseline tracker uses the backbone network to extract features, and then uses the cross-correlation operation to complete the similarity calculation of the two images. Through experiments, it is found that only the features proposed by the backbone network affect the tracking performance. Therefore, when out of view, similar objects and illumination changes are encountered, the tracker is vulnerable to interference. Therefore, after the backbone network, it is particularly important to improve the correlation of remote pixel pairs and the enhancement of target feature information. Therefore, UAVs are easily disturbed by occlusion, similar objects and illumination changes. It is particularly important to improve the correlation between the template frame and detection frame and highlight the target features. As shown in Figure 1, this study uses a red dotted box to indicate that it has the same structure as the baseline tracker but different parameters. In the template frame and detection frame, input images are cut and filled to sizes of 127×127×3 and 303×303×3 respectively. Then, the tensor sizes obtained by the feature extraction networks are 6×6×512 and 28×28×512 respectively. The backbone network parameters are shown in Table 1. Then, they are fed into the AIFM and FDCM designed in this paper respectively, and the association between the adaptive learning template frame target and detection frame target is carried out. After 3×3 convolutions, a deep cross-correlation calculation is carried out. There are two kinds of task heads in the network, which are mainly used to complete the classification of the target and background and the regression task of bounding boxes. Finally, the bounding box with the highest response score is selected as the tracking result through non-maximum suppression (NMS).

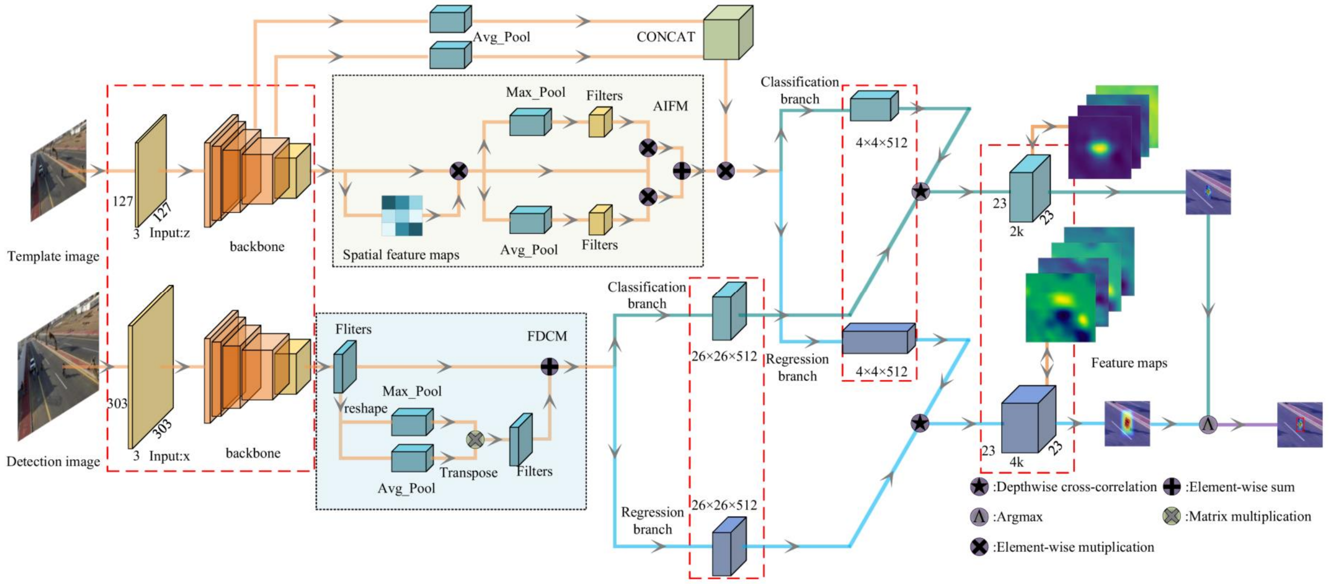

Figure 1. *UAV tracker network structure is designed in this paper.*

Table 1. *Backbone network parameters.*

layers	stride	kernel	channel	template	detection
Input	-	-	-	127×127	303×303
Conv1	2	11×11	96	59×59	147×147
MaxPool1	2	3×3	96	29×29	73×73
Conv2	1	5×5	256	25×25	69×69
MaxPool2	2	3×3	256	12×12	34×34
Conv3	1	3×3	768	10×10	32×32
Conv4	1	3×3	768	8×8	30×30
Conv5	1	3×3	512	6×6	28×28

3.2. Template Frame Feature Extraction

The attention information fusion module (AIFM) can automatically increase the weight of target-related information in spatial and channel dimensions by extending the perceptual field in convolutional neural networks and fusing features at different levels to effectively build a feature cascade of remote pixel pairs. Two one-dimensional kernels are used to build local context models in the vertical and horizontal directions respectively to capture the target location information. Then, the local cross-channel interaction strategy is adopted to consider the correlation between each channel and the adjacent K channels. Numerous experimental results show that reducing the dimensionality in the template frame has side effects on the prediction results of tracking.

The AIFM averages and slides the eigenvalues of each row and each column along the vertical and horizontal directions of the window, to capture more abundant spatial semantic information and effectively suppress the background information. The size of the pooling window is set to (1, *W*) and (*H*, 1). The input tensor is $a \in R^{H \times W}$. Vertical and horizontal directions can be calculated as:

$$y_i^h = \frac{1}{W} \sum_{0 \le j \le W} a_{i,j}, y^h \in R^H \quad (1)$$

$$y_i^v = \frac{1}{H} \sum_{0 \le j \le W} a_{i,j}, y^v \in R^V \quad (2)$$

The input tensor of AIFM is $x \in R^{C \times H \times W}$, C represents the number of channels, and H and W represent the height and width, respectively. Then, one-dimensional convolution with kernel 3 is used to modulate the feature information of the adjacent position and the current position to obtain the spatial information of the remote global context. $y^h \in R^{C \times H}$ and $y^v \in R^{C \times W}$ can be gained respectively. By fusing y^h and y^v, $y \in R^{C \times H \times W}$ can be gained. The mathematical expression is:

$$y_{c,i,j} = Add(y_{c,i}^h, y_{c,j}^v) \quad (3)$$

where $Add(.,.)$ denotes element-wise sum among feature maps. Next, $z \in R^{C \times H \times W}$ can be gained, and the mathematical expression is as follows:

$$z = Scale(x, \sigma(f(y))) \quad (4)$$

f describes 1×1 convolution, $Scale(.,.)$ describes elementwise mutiplication. σ describes the sigmoid function. The pooling window sizes are given as 16×16 and 12×12 respectively, and batch normalization (BN) and ReLU operations are performed after each layer of convolution to further boost the feature extraction capability. Channel C can be expressed as:

$$C = [C_1, C_2, C_3, \cdots, C_{512}] \quad (5)$$

Next, the obtained z is further processed using global maximum pooling (GMP) and global average pooling (GAP) to obtain rich context information. In this tracker, the band matrix W_k is used to express the learned channel weight information, so that the number of cross-channel interactions can be implemented. W_k can be formulated as:

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \quad (6)$$

In this tracker, only the k weight information adjacent to z_i is considered, which can be formulated as:

$$w_i = \sigma\left(\sum_{j=1}^{k} w^j z_i^j\right), z_i^j \in \Omega_i^k \quad (7)$$

where Ω_i^k denotes the complete set of k channels adjoining to z_i, and w^j indicates the shared weight information. Then, the strategy can be realized by fast one-dimensional convolution (represented as $C1D$) with convolution kernel size of k. Mathematical, it can be expressed as:

$$w = \sigma(C1D_k(z)) \quad (8)$$

Then, this study can obtain w_{gmp} and w_{gap} respectively. Mathematical, they can be expressed as:

$$w_{gmp} = \sigma\left(C1D_k\left(z\right)\right) \quad (9)$$

$$w_{gap} = \sigma\left(C1D_k\left(z\right)\right) \quad (10)$$

Since the number of channels in a network is usually a multiple of 2, there is a linear mapping between C and k, and the mathematical expression of ψ is:

$$k = \psi\left(C\right) = \left|\frac{\log_2\left(C\right)}{2} + \frac{1}{2}\right|_{odd} \quad (11)$$

where $\left|p\right|_{odd}$ denotes the nearest odd number closest to p, and then $z_{gmp} \in R^{C\times H\times W}$ and $z_{gap} \in R^{C\times H\times W}$ are obtained, which can be expressed as:

$$z_{gmp} = Scale(w_{gmp}, z) \quad (12)$$

$$z_{gap} = Scale(w_{gap}, z) \quad (13)$$

$r \in R^{C\times H\times W}$ is obtained after feature fusion, i.e.:

$$r = Add(z_{gmp}, z_{gap}) \quad (14)$$

After obtaining r, the features are further processed by BN and ReLU. Then, one-dimensional convolution and multilayer fusion are performed.

Next, this study introduces CycleMLP to make full use of the shallow feature information and capture more feature information. The input tensors are $x_1 \in R^{C_1\times H_1\times W_1}$ and $x_2 \in R^{C_2\times H_2\times W_2}$, which represent the feature maps of the third and fourth layers of the backbone network respectively. Then, after GAP calculation, this study concatenates the feature information, and obtain the prediction result through a multilayer perceptron (MLP). After reshaping, elementwise mutiplication is employed to strengthen the aggregation of shallow feature information. Thus, $l \in R^{C\times H\times W}$ is obtained, i.e.:

$$l = Scale\left(x_1, \left(MLP\left(cat\left(AvgPool(x_2), AvgPool(x_1)\right)\right)\right)^S\right) \quad (15)$$

where s denotes reshape and cat denotes concat. $AvgPool$ denotes GAP calculation. MLP represents two fully connected layers with BN and ReLU in each layer. After CycleMLP, two layers of convolution with a convolution kernel size of 3 are used to complete the calculation of deep features. Again with the deep feature fusion, highlighting the target information, the obtained $g \in R^{C\times H\times W}$ mathematical formula can be expressed as:

$$g = Scale\left(W_{G_2}\left(\mathrm{Re}LU\left(LN\left(W_{G_1}\left(\mathrm{Re}LU\left(LN\left(l\right)\right)\right)\right)\right)\right), r\right) \quad (16)$$

where LN represents the batch normalization, and W_{G_1} and W_{G_2} denote the weight parameters learned by the two layers of convolution, respectively. Finally, the learned weight vector can be obtained, and the final channel $\overline{C}$ can be expressed as:

$$\overline{C} = \alpha \cdot C = [\overline{C}_1, \overline{C}_2, \overline{C}_3, \cdots, \overline{C}_{256}] \quad (17)$$

3.3. Detection Frame Feature Extraction

Because the detection frame image is constantly changing, it is particularly important to enhance the feature extraction ability of the detection frame dynamic image target information. If the correlation between the extracted feature information and target information is small, the performance of the tracker will be seriously affected. This study design a feature deep convolution module (FDCM) that can adaptively learn and weight the target feature information in the detection frame, to improve the target positioning ability of the tracker.

The input tensor is $x \in R^{C \times H \times W}$, where C represents the number of channels and H and W represent the height and width, respectively. First, after convolution with a kernel size of 1×1, the feature information is processed by batch normalization and an activation function. A large number of experiments show that too many detection frame channels will carry a large amount of feature information irrelevant to the target. Therefore, compressing the number of channels is an effective method to learn target features, and also reduces the number of parameters. In the experiment, r is the compression channel ratio, and the size is 8. Then the obtained feature information is reshaped to obtain $\{B, D\} \in R^{C/r \times N}$ where N represents $H \times W$, i.e.:

$$B = \operatorname{Re}LU\left(LN\left(W_{V1} \times x\right)\right) \tag{18}$$

$$D = \operatorname{Re}LU\left(LN\left(W_{V1} \times x\right)\right) \tag{19}$$

LN represents the batch normalization operation, and W_{V1} represents the weight information learned by deep convolution. B and D learn the feature information of the target through GAP and GMP respectively, and complete the matrix multiplication operation. After taking the maximum value of the obtained output feature and completing the reshaping, the normalized probability distribution of the target feature is obtained through the softmax function, which can adaptively learn the spatial position information of the target and output $y \in R^{C \times H \times W}$ can be expressed as:

$$y = soft\max\left(\max\left(M\operatorname{ax}Pool\left(B\right) \times AvgPool\left(D\right)^{T}\right)^{S}\right) \tag{20}$$

where, T represents transpose and s represents reshape. Then, this study extend the channel through convolution and allocate various weight information to learn the relevant features of the target, to realize the information fusion of the features and attain the final output Z, i.e.:

$$Z = \operatorname{Re}LU\left(LN\left(Add\left(W_{V2} \times y, x\right)\right)\right) \tag{21}$$

Where W_{V2} represents the model weight learned by extended channel convolution.

4. Experiments

4.1. Experimental Environment and Datasets

The UAV tracking algorithm in this paper is based on PyTorch1.4 in the Linux system. The sole GPU is GTX 2060Super 8G, and the CPU is Intel Core i7-9700F @ 3.00GHz.

The ILSVRC2017_VID dataset and Youtube-BB dataset are employed for model training, including 45800 video sequences with real labels, containing more than one million frames. The algorithm is tested on UAV123 [26] dataset and UAV20L [26] dataset. The UAV123 dataset contains 123 video sequences with 12 attribute variations, including scale variation, aspect ratio change, low resolution, fast motion, full occlusion, partial occlusion, out-of-view, background clutter, illumination variation, viewpoint change, camera motion, and similar object. It is also the largest UAVs tracking dataset at present. The UAV20L dataset contains 20 video sequences with 12 attribute variations and is a subset of the UAV123 dataset. It is primarily utilized to evaluate the UAVs long-term tracking problems. DTB70 contains 70 video sequences with 11 attribute variations, which is often accompanied by serious camera jitter during shooting.

4.2. Evaluation Metrics

The one-pass evaluation (OPE) metric is utilized to evaluate the tracking performance of UAVs, including the success rate and precision. The success rate is the ratio of the number of bounding boxes to the number of real bounding boxes in the previous frame that are greater than a set threshold, and can be expressed as an area under the curve (AUC) success curve score. Here, intersection over union (IOU) is the most direct calculation indicator. The precision is evaluated by the center location error (CLE) between the bounding box and the real bounding box. The precision plot is drawn by the percentage of the bounding boxes whose CLE is less than the preset threshold in the total bounding boxes of the previous frame. In the experiment, the threshold value is defined as 20 pixels. The distance precision rate (DP) is obtained from the value of the precision curve. There were 240000 iterations during the training. Stochastic gradient descent (SGD) is employed for the gradient update with momentum = 0.9. This study set the dynamic learning rate to be initialized to 3×10^{-2} and reduced it to 10^{-5}. When the response value of the complete intersection over union is greater than 0.6, the predicted bounding box is a positive sample, and less than 0.3 is regarded as a negative sample.

4.3. Experimental Results

1) Results on the UAV123 dataset: In this experiment, this study employ the SiamRPN [8] tracker as the baseline tracker, and 17 most advanced trackers are selected to prove the effectiveness of the proposed UAVs visual tracker SiamDFT++, including: HiFT [17], AutoTrack [20], SiamRPN [8], SiamFC [4], MCCT [21] (multi-cue correlation filters), MCPF [22] (multi-task correlation particle filter), ECO_HC [17], DeepSTRCF [7], SRDCF [23] (learning spatially regularized correlation filters), ARCF [28] (learning aberrance repressed correlation filters), UDT [9], BACF [6], Staple [3], TADT

[10], SAMF [14], and DSST [5] and KCF [2]. As shown in Table 2:

Table 2. Quantification evaluation on UAV123.

Trackers	Venue	DP	AUC
SAMF	ECCV2014	0.593	0.396
DSST	BMVC2014	0.586	0.356
KCF	TPAMI2015	0.523	0.331
SRDCF	ICCV2015	0.676	0.463
Staple	CVPR2016	0.595	0.409
SiamFC	ECCV2016	0.696	0.480
MCPF	CVPR2017	0.718	0.473
BACF	ICCV2017	0.660	0.459
ECO_HC	CVPR2017	0.710	0.496
DeepSTRCF	CVPR2018	0.705	0.508
MCCT	CVPR2018	0.734	0.507
SiamRPN	CVPR2018	0.749	0.528
TADT	CVPR2019	0.727	0.520
ARCF	ICCV2019	0.671	0.468
UDT	CVPR2019	0.668	0.477
AutoTrack	CVPR2020	0.689	0.472
HiFT	ICCV2021	0.787	0.589
SiamDFT++		0.811	0.592

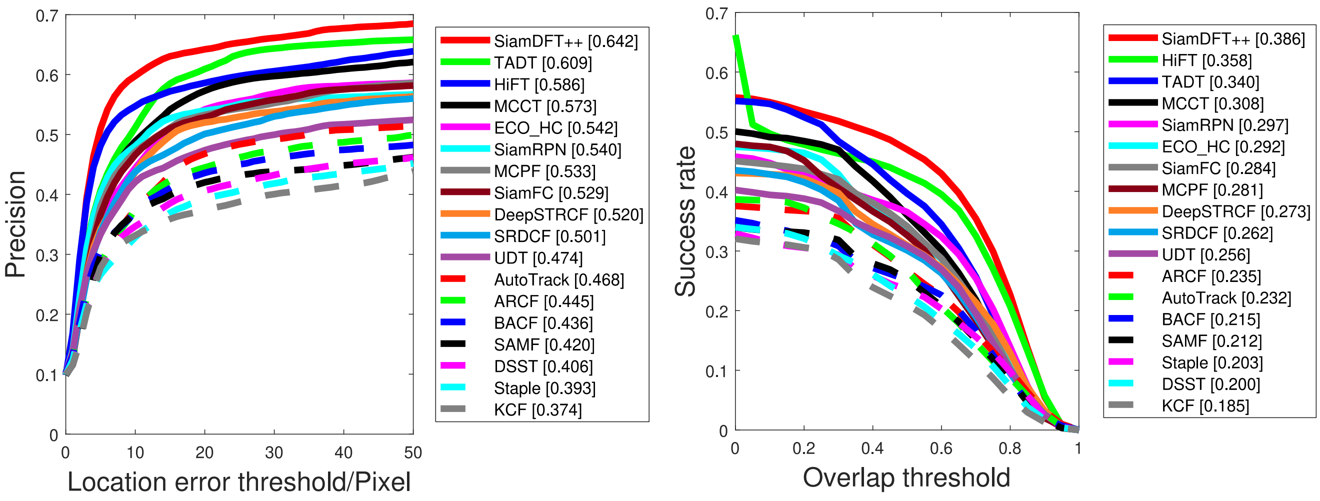

Figure 2. Full occlusion of similar target scenarios.

Compared with the latest tracker, the tracker has achieved the distance precision rate (DP) of 81.1% and an AUC score of 59.2%, ranking first among all trackers. Compared with the latest HiFT using a transformer structure, the positioning accuracy of the target is improved by 3.1%. Compared with the benchmark tracker, the overall precision is improved by 8.3%, and the success rate is improved by 12.12%. As shown in Figure 2. The tracker shows the effectiveness of the tracker under the target full occlusion attribute, and the DP and AUC are improved by 18.9% and 30.0%, respectively.

As shown in Figure 3. Due to the large viewing angle range and the problem of similar targets often encountered in UAVs visual tracking, SiamDFT++ can also track targets effectively. The tracking accuracy and success rate are 77.9% and 55.9%, respectively, which effectively proves the robustness of the tracker in this paper. It can be seen that only using manual features (such as KCF, SAMF, ARCF) or lightweight models (such as SiamFC, SiamRPN, TADT) leads to poor robustness of the tracker. Even though HiFT uses the transformer structure to enhance the dependency between global information, the local information is poor, resulting in only 51.4% of the AUC of similar targets.

In fact, SiamDFT++ shows a good tracking effect under various attributes, as shown in Table 3. Compared with traditional visual tracking, UAVs visual tracking produces strong camera jitter and is more vulnerable to illumination. Therefore, the tracker considers the ability of target feature extraction by the template frame and detection frame at the same time, and establishes an appearance feature model that can adaptively learn the target. SiamDFT++ can effectively predict and locate targets and better adapt to UAVs visual tracking scenes.

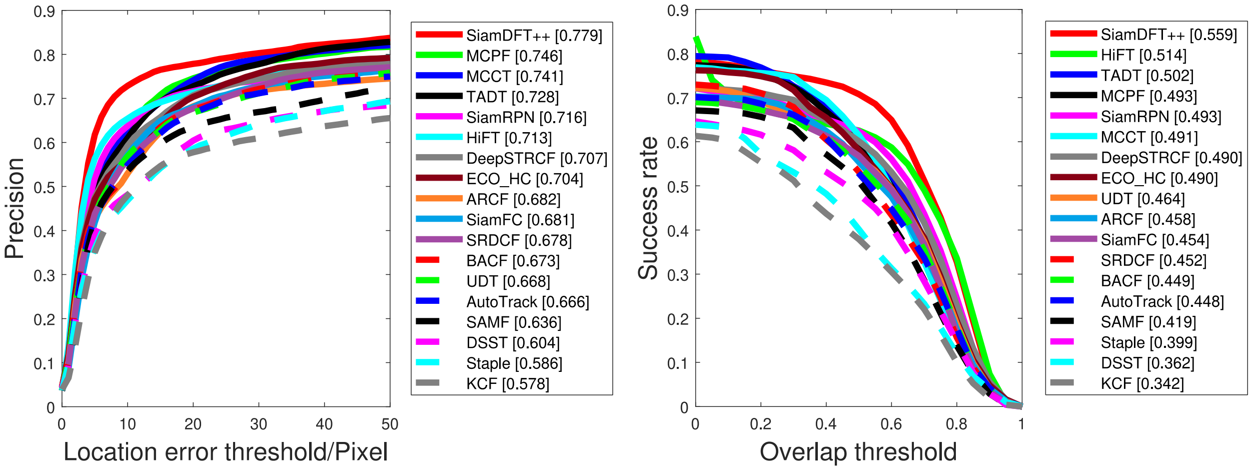

***Figure 3**. Similar object of similar target scenarios.*

As shown in Figure 4, this study selected truck1, car18 and car6_2 scenarios to verify the robustness of the SiamDFT++ tracker. It can be seen from the truck1 scene that SiamDFT++ can still locate the target well when the tracker is affected by background interference, target occlusion, similar objects and illumination variation. From the car18 scene, it can be seen that the red car is moving rapidly, and the tracker proposed can still locate the target quickly and accurately. It can be seen from the car6_2 scene that the tracker can still accurately locate the target in the scene with multiple challenges when tracking the vehicle at high altitude, accompanied by camera jitter and partial occlusion, as well as the scale variation and viewpoint change.

***Table 3**. Comparisons of algorithms for camera motion and illumination variation.*

scence	Camera Motion Illumination Variation			
	DP	AUC	DP	AUC
SAMF	0.561	0.381	0.478	0.312
DSST	0.520	0.322	0.524	0.307
KCF	0.483	0.310	0.418	0.270
SRDCF	0.627	0.439	0.600	0.395
Staple	0.544	0.386	0.498	0.362
SiamFC	0.684	0.482	0.603	0.391
MCPF	0.700	0.463	0.659	0.424
BACF	0.639	0.450	0.525	0.356
ECO_HC	0.676	0.476	0.628	0.407
DeepSTRCF	0.696	0.509	0.664	0.444
MCCT	0.720	0.508	0.704	0.466
SiamRPN	0.750	0.537	0.665	0.456
TADT	0.723	0.518	0.669	0.462
ARCF	0.647	0.455	0.595	0.392
UDT	0.654	0.467	0.599	0.401
AutoTrack	0.658	0.458	0.617	0.396
HiFT	0.799	0.600	0.700	0.502
SiamDFT++	0.831	0.615	0.805	0.573

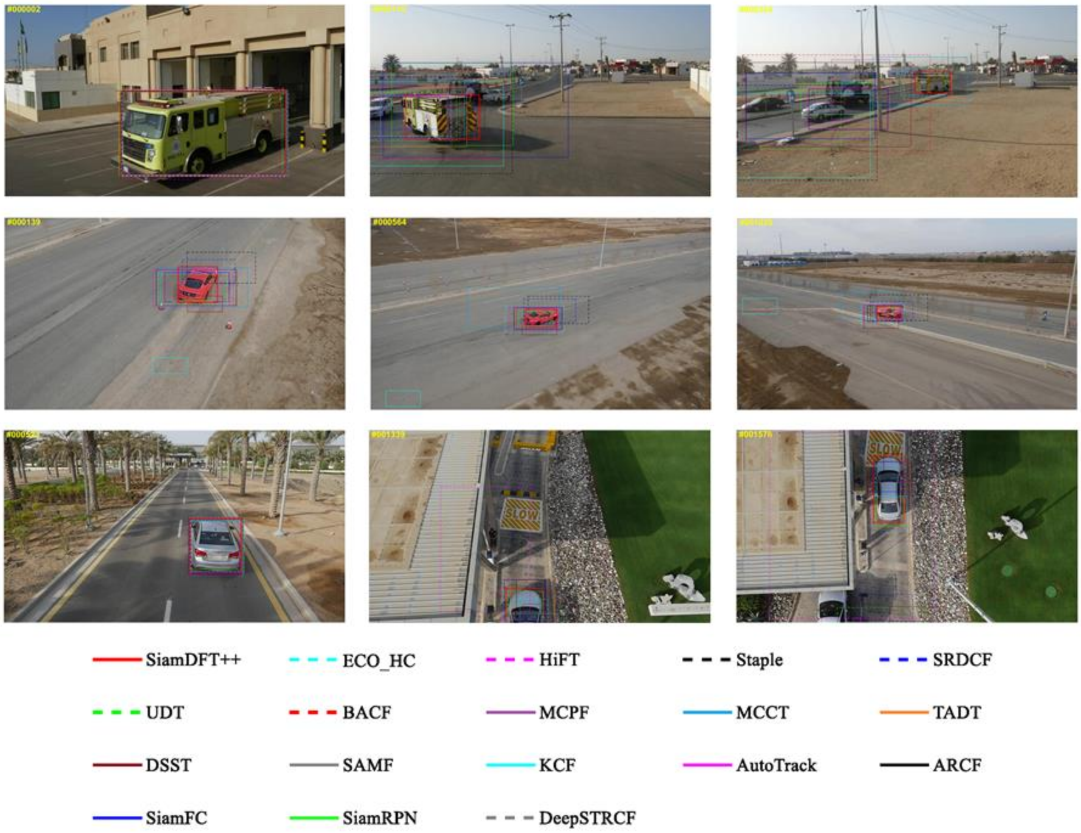

***Figure 4**. Real scenes tracking on UAV123 dataset.*

2) Results on the UAV20L dataset: The UAV20L dataset focuses on the UAVs long-time tracking problem. Therefore, in order to verify the effectiveness of the SiamDFT++ tracker, this study test the tracker on the UAV20L dataset and compare it with 19 kinds of advanced trackers, including: AutoTrack, DaSiamRPN (distractor-aware siamese networks), SiamRPN, DSiam (dynamic siamese network), SiamFC, MCCT, MCPF, ECO_HC, DeepSTRCF, SRDCF, ARCF, HiFT, UDT+, BACF, Staple, TADT, SAMF, CCOT [15] (continuous convolution operators), and KCF. As shown in Table 4, the experimental results show that the UAVs visual tracker designed in this paper is still reliable in long term tracking. The DP and AUC are 72.3% and 54.6%, respectively, and the DP and AUC are increased by 15.50% and 18.18%, respectively.

***Table 4**. Quantification evaluation on UAV20L.*

Trackers	DP	AUC	Trackers	DP	AUC
SAMF	0.470	0.326	DeepSTRCF	0.588	0.443
CCOT	0.561	0.395	MCCT	0.605	0.407
KCF	0.311	0.196	SiamRPN	0.626	0.462
SRDCF	0.507	0.343	DaSiamRPN	0.665	0.465
Staple	0.455	0.331	ARCF	0.544	0.381
SiamFC	0.613	0.399	UDT+	0.585	0.401
MCPF	0.586	0.370	AutoTrack	0.512	0.349
BACF	0.584	0.415	TADT	0.609	0.459
ECO_HC	0.522	0.387	HiFT	0.763	0.566
DSiam	0.603	0.391	SiamDFT++	0.723	0.546

3) Results on the DTB70 dataset: this study tested the tracker on the DTB70 dataset and compared it with 17 kinds of most advanced trackers, including: HiFT [17], Auto Track [20], UDT+ [9], ARCF [28], UDT [9], TADT [10], CCOT [15], SiamRPN [8], CFNet conv2 [23], MCCT [21], DeepSTRCF [7], ECO_gpu [16], BACF [6], MCPF [22], SRDCF [23], KCF [2], and SAMF [14]. As shown in Table 5, the experimental results show that the UAVs visual tracker designed in this paper is still reliable in the special design of camera jitter. The DP and AUC are increased by 10.54% and 16.23%, respectively. ECO_gpu uses multifeature fusion to extract target information, SiamRPN uses the regression fine-tuning mechanism of bounding box, and UDT proposes multi-frame verification strategy to improve tracking performance. However, it can be seen from the data that these trackers are redundant in feature extraction, resulting in low environmental adaptability.

Table 5. Quantification evaluation on DTB70.

Trackers	DP	AUC	Trackers	DP	AUC
SAMF	0.519	0.340	SiamRPN	0.721	0.499
KCF	0.468	0.280	CCOT	0.769	0.517
SRDCF	0.512	0.363	TADT	0.693	0.464
MCPF	0.664	0.433	ARCF	0.694	0.472
BACF	0.590	0.402	UDT	0.602	0.422
ECO_gpu	0.722	0.502	UDT+	0.658	0.462
DeepSTRCF	0.734	0.506	AutoTrack	0.717	0.479
MCCT	0.725	0.484	HiFT	0.802	0.594
CFNet_conv2	0.616	0.415	SiamDFT++	0.797	0.580

Table 6. The analysis table of method validity is on got10k benchmark.

Method	UAV123 [26]		UAV20L [26]		DTB70 [27]	
	DP	AUC	DP	AUC	DP	AUC
BT	0.734	0.532	0.746	0.498	0.719	0.499
BT + F	0.757	0.569	0.739	0.520	0.744	0.534
BT + F + AIFM	0.773	0.579	0.653	0.519	0.757	0.543
BT + F + FDCM	0.780	0.582	0.662	0.523	0.771	0.558
SiamDFT	0.786	0.593	0.677	0.548	0.783	0.560
SiamDFT++	0.798	0.601	0.695	0.555	0.797	0.580

4) Ablation experiments: this study design ablation experiments on the got10k benchmark. As shown in Table 6, this study use BT to represent the baseline tracker, and use F to represent the enhancement of the width of the backbone network, the introduction of the complete intersection over union method and deep cross-correlation operation. Before using shallow features, this study referred to the tracker as SiamDFT. It can be seen from the data that SiamDFT++ has good robustness in both short-term and long-term tracking scenarios, which proves the effectiveness of this algorithm. At the same time, the tracker runs with 66FPS (frames per second), which has good real time performance.

As shown in Figure 5, this study use DJI Maciv Air2 to test the SiamDFT++ in a college sports ground, and the flight altitude is 15 meters. this study also display the heatmap and response map in the tracking process, and the tracker can always locate at the center of the target.

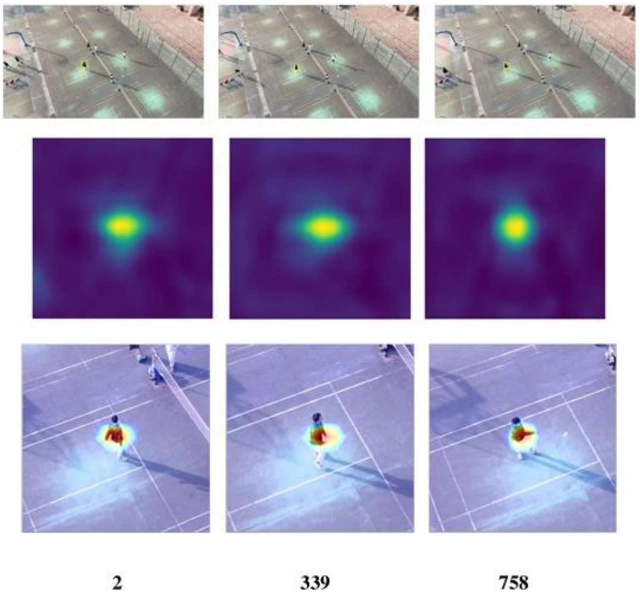

Figure 5*. Real scene tracking on the UAV platform.*

5. Conclusions

In this paper, aiming at the influence of object occlusion, camera jitter and similar objects in the process of UAVs visual tracking, a UAVs visual tracker (SiamDFT++) is designed to improve the tracking performance of a siamese network with regression calculation, introduce a deep cross-correlation operation to strengthen the accuracy of similarity calculation and improve the number of channels to increase the appearance feature information of the target. The calculation method of complete intersection over union is introduced to complete the calculation of the target frame, the attention information fusion model is proposed, and the shallow features are fully used to improve the extraction ability of the template frame target. The feature deep convolution network is designed to adapt to the learning and detection of the appearance information of the frame target, so as to effectively improve the visual tracking performance of UAVs. In the future, it is considered to improve the detection ability of the tracker for the target features of the first frame will be improved. When the target drifts or loses, it will be redetected without affecting the real-time tracking, so as to improve the robustness of the UAVs visual tracker.

Abbreviations

UAVs	Unmanned Aerial Vehicles

Author Contributions

Shuduo Zhao: Conceptualization, Methodology, Funding acquisition

Yunsheng Chen: Writing – original draft, Formal Analysis

Shuaidong Yang: Software, Validation

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analysed in this study. If anyone is interested in our Python simulation, please contact the first author, we are glad to provide the Python code.

Funding

This work is supported by the Special Project of Science and Technology Strategic Cooperation between Nanchong City and Southwest Petroleum University (23XNSYSX0022).

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Yang. S, Xu. J, Chen. H, et al, High-performance UAVs visual tracking using deep convolutional feature, Neural Computing and Applications, 2022, pp. 13539-13558. https://doi.org/10.1007/s00521-022-07181-w

[2] Henriques. J. F, Caseiro. R, Martins. P, et al, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern. Anal. vol. 37, no.3, pp. 583-596, Mar. 2015. https://doi.org/10.1109/TPAMI.2014.2345390

[3] Bertinetto. L, Valmadre. J, Golodetz. S, et al, Staple: Complementary learners for real-time tracking, presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, Jun.2021, pp. 1401-1409. https://doi.org/10.48550/arXiv.1512.01355

[4] Bertinetto. L, Valmadre. J, Henriques. J. F, et al, Fully-convolutional siamese networks for object tracking, European conference on computer vision, Nov. 2016, pp. 850-865. https://doi.org/10.48550/arXiv.1606.09549

[5] Danelljan. M, Hager. G, Khan. F, et al, Accurate scale estimation for robust visual tracking, presented at British Machine Vision Conference, Sep. 2014, pp. 1-5. https://doi.org/10.5244/c.28.65

[6] Kiani. Galoogahi. H, Fagg. A, Lucey. S, Learning background aware correlation filters for visual tracking, presented at Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Oct.2017, pp. 1135-1143. https://doi.org/10.1109/ICCV.2017.129

[7] Li. F, Tian. C, Zuo. W, et al, Learning spatial-temporal regularized correlation filters for visual tracking, presented at Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun.2018, pp. 4904-4913. https://doi.org/10.1109/CVPR.2018.00515

[8] Li. B, Yan. J, Wu. W, et al, High performance visual tracking with siamese region proposal network, presented at Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 8971-8980. https://doi.org/10.1109/CVPR.2018.00935

[9] Wang. N, Song. Y, Ma. C, et al, Unsupervised Deep Representation Learning for Real-Time Tracking, International Journal of Computer Vision, Jun. 2021, pp. 400-418. https://doi.org/10.1007/s11263-020-01357-4

[10] Li. X, Ma. C, Wu. B, et al, Target-Aware Deep Tracking, presented at Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2019, pp. 1369-1378. https://doi.org/10.48550/arXiv.1904.01772

[11] Li. B, Wu. W, Wang. Q, et al, Siamrpn++: Evolution of siamese visual tracking with very deep networks, presented at Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun.2019, pp. 4282-4291. https://doi.org/10.48550/arXiv.1812.11703

[12] Hou. Q, Zhang. L, Cheng. M. M, et al, Strip pooling: Rethinking spatial pooling for scene parsing, presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2020, pp. 4003–4012. https://doi.org/10.48550/arXiv.2003.13328

[13] Wang. Q, Wu. B, Zhu. P, et al, ECA-Net: Efficient channel attention for deep convolutional neural networks, presented at 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2020. https://doi.org/10.48550/arXiv.1910.03151

[14] Li. Y, Zhu. J, A scale adaptive kernel correlation filter tracker with feature integration, presented at European conference on computer vision, Mar. 2015, pp. 254-265. https://doi.org/10.1007/978-3-319-16181-5_18

[15] Danelljan. M, Robinson. A, Khan. F. S, et al, Beyond correlation filters: Learning continuous convolution operators for visual tracking, presented at European conference on computer vision, Sep. 2016, pp.472-488. https://doi.org/10.1007/978-3-319-46454-1_29

[16] Danelljan. M, Bhat. G, Shahbaz. KhanF, et al, Eco: Efficient convolution operators for tracking, presented at Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jul. 2017, pp. 6638-6646. https://doi.org/10.48550/arXiv.1611.09224

[17] Cao. Z, Fu. C, Ye. J, et al, HiFT: Hierarchical Feature Transformer for Aerial Tracking, presented at Proceedings of the IEEE/CVF International Conference on Computer Vision, Oct. 2021, pp. 15457-15466. https://doi.org/10.1109/ICCV48922.2021.01517

[18] Chen. S, Xie. E, Ge. C, et al, CycleMLP: A MLP-like Architecture for Dense Visual Predictions, IEEE Transactions on Pattern Analysis and Machine Intelligence, Nov. 2023, pp. 1-17. https://doi.org/10.1109/TPAMI.2023.3303397

[19] Zheng. Z, Wang. P, Liu. W, et al, Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression, presented at Proceedings of the AAAI Conference on Artificial Intelligence, Feb. 2020, pp. 12993-13000. https://doi.org/10.48550/arXiv.1911.08287

[20] Li. Y, Fu. C, Ding. F, et al, Auto Track: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization, presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2020, pp. 11923-11932. https://doi.org/10.48550/arXiv.2003.12949

[21] Wang. N, Zhou. W, Tian. Q, et al, Multi-cue correlation filters for robust visual tracking, presented at Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 4844-4853. https://doi.org/10.1109/CVPR.2018.00509

[22] Zhang. T, Xu. C, Yang. M. H, Multi-task correlation particle filter for robust object tracking, presented at Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jul. 2017, pp. 4335-4343. https://doi.org/10.1109/CVPR.2017.512

[23] Danelljan. M, Hager. G, Shahbaz. Khan. F, et al, Learning spatially regularized correlation filters for visual tracking, presented at Proceedings of the IEEE International Conference on Computer Vision, Dec.2015, pp. 4310-4318. https://doi.org/10.48550/arXiv.1608.05571

[24] Yang. K, He. Z, Pei. W, et al, Siam Corners: Siamese corner networks for visual tracking, IEEE Transactions on Multimedia, vol. 24, pp.1956-1967, Apr. 2021. https://doi.org/10.1109/TMM.2021.3074239

[25] Xu. Y, Wang. Z, Li. Z, et al, Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines, presented at Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no.07, pp. 12549-12556, Apr. 2020. https://doi.org/10.48550/arXiv.1911.06188

[26] Mueller. M, Smith. N, Ghanem. B, A benchmark and simulator for uav tracking, presented at European conference on computer vision, Sep. 2016, pp. 445-461. https://doi.org/10.1007/978-3-319-46448-0_27

[27] Li. S, Yeung. DY, Visual object tracking for unmanned aerial vehicles: a benchmark and new motion models, presented at Proceedings of the AAAI conference on artificial intelligence, Feb. 2017, pp. 4140-4146.

[28] Huang. Z, Fu. C, Li. Y, et al, Learning aberrance repressed correlation filters for real-time UAV tracking, presented at Proceedings of the IEEE/CVF International Conference on Computer Vision, Oct. 2019, pp. 2891-2900. https://doi.org/10.48550/arXiv.1908.02231

[29] Yang. S, Chen. H, Xu. F, et al, High-performance UAVs visual tracking based on siamese network, The Visual Computer, 2021. https://doi.org/10.1007/s00371-021-02271-7

Biography

Shuduo Zhao is an associate professor at the School of Electrical Information, Southwest Petroleum University, where she obtained a master's degree. Chengdu, China, In 2009. She focuses on electronic technology applications, neural networks, and machine learning. She teaches students basic programming skills and supervises papers to shape future electronic engineers.

Yunsheng Chen: a senior engineer at the School of Science, Southwest Petroleum University, earned his Bachelor of Science degree there, in Chengdu, China, in 2007. He specializes in electronic technology applications. Chen has guided students in numerous electronic design competitions, earning many awards.

Shuaidong Yang: from Chongqing, China, is currently pursuing his doctoral studies at the School of Automation, Chongqing University. He has earned both a Bachelor's and a Master's degree in Engineering from Southwest Petroleum University. Building on his expertise in image processing, he will delve into reinforcement learning to develop intelligent machine learning algorithms.