drones

MDPI

Article

Anti-Occlusion UAV Tracking Algorithm with a Low-Altitude Complex Background by Integrating Attention Mechanism

Chuanyun Wang [1,*], Zhongrui Shi [2], Linlin Meng [1], Jingjing Wang [3], Tian Wang [4], Qian Gao [1] and Ershen Wang [5]

[1] College of Artificial Intelligence, Shenyang Aerospace University, Shenyang 110136, China; menglinlin@stu.sau.edu.cn (L.M.); gaoqian@buaa.edu.cn (Q.G.)
[2] School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China; shizhongrui@stu.sau.edu.cn
[3] China Academic of Electronics and Information Technology, Beijing 100041, China; wangjingjing@cetccloud.com
[4] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China; wangtian@buaa.edu.cn
[5] School of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang 110136, China; wes2016@sau.edu.cn
* Correspondence: wangcy0301@sau.edu.cn

Abstract: In recent years, the increasing number of unmanned aerial vehicles (UAVs) in the low-altitude airspace have not only brought convenience to people's work and life, but also great threats and challenges. In the process of UAV detection and tracking, there are common problems such as target deformation, target occlusion, and targets being submerged by complex background clutter. This paper proposes an anti-occlusion UAV tracking algorithm for low-altitude complex backgrounds by integrating an attention mechanism that mainly solves the problems of complex backgrounds and occlusion when tracking UAVs. First, extracted features are enhanced by using the SeNet attention mechanism. Second, the occlusion-sensing module is used to judge whether the target is occluded. If the target is not occluded, tracking continues. Otherwise, the LSTM trajectory prediction network is used to predict the UAV position of subsequent frames by using the UAV flight trajectory before occlusion. This study was verified on the OTB-100, GOT-10k and integrated UAV datasets. The accuracy and success rate of integrated UAV datasets were 79% and 50.5% respectively, which were 10.6% and 4.9% higher than those of the SiamCAM algorithm. Experimental results show that the algorithm could robustly track a small UAV in a low-altitude complex background.

Keywords: unmanned aerial vehicle; target tracking; attention mechanism; anti-occlusion; location prediction

Citation: Wang, C.; Shi, Z.; Meng, L.; Wang, J.; Wang, T.; Gao, Q.; Wang, E. Anti-Occlusion UAV Tracking Algorithm with a Low-Altitude Complex Background by Integrating Attention Mechanism. *Drones* **2022**, *6*, 149. https://doi.org/10.3390/drones6060149

Academic Editors: Daobo Wang and Zain Anwar Ali

Received: 11 May 2022
Accepted: 14 June 2022
Published: 16 June 2022

1. Introduction

In recent years, with the rapid development of the UAV industry and the continuous improvement of artificial intelligence, UAVs have been widely used in public security, disaster relief, photogrammetry, news broadcasts, travel, and other fields, bringing great convenience to production and social life. However, the increasing number of UAVs in low-altitude airspace and the frequent occurrence of various illegal flight incidents have brought great threats and challenges to aviation flight, security, confidentiality protection and privacy protection [1].

In order to detect UAV in the low-altitude airspace as early and as far as possible using computer vision, it is often necessary to implement the long-distance detection and tracking of UAVs [2], which cause small imaging sizes and weak signals [3]. At the same time, the flight altitude of UAV in low-altitude airspace is very low, often only dozens to hundreds of meters, and the surrounding environment of this altitude is relatively complex, such as trees, buildings, and walls, which may lead to the visual tracking of UAV being interfered

by strong clutter, occlusion, and other factors, resulting in tracking drift and loss. Therefore, it is important and urgent to find a robust tracking algorithm against background clutter interference and occlusion for UAV tracking in low-altitude airspace.

On the basis of combining existing UAV visual tracking technology and referring to the network structure of ATOM [4], this paper proposes an anti-occlusion target tracking algorithm by integrating the SeNet [5] attention mechanism to solve the complex background and occlusion problems during tracking, which achieved good performance. First, the SeNet attention mechanism was introduced into the original feature extraction network to enhance the extracted features, which effectively improved the performance of subsequent tracking process. Second, an occlusion-sensing model was designed to judge the state of the target. Lastly, the LSTM [6] trajectory prediction network was used to predict the UAV position according to the target state. This study was verified on the OTB-100 [7], GOT-10k [8] and integrated UAV datasets. Experimental results show that the proposed algorithm could effectively reduce the influence of low-altitude complex environments on the target and robustly track a UAV.

The main contributions of this paper are:

1. In order to solve the problem of UAVs in the low-altitude airspace being easy to be submerged in complex background clutter, the SeNet attention mechanism was used in the backbone to improve the correlation between feature channels, enhance the feature of the target, and reduce the influence of background clutter.

2. In order to solve the occlusion problem in low-altitude airspace during flight, an occlusion judgment mechanism is proposed to judge whether the target is occluded.

3. When the target is occluded, the LSTM trajectory prediction network is used to predict the flight trajectory of the aircraft, so as to achieve robust tracking and stop the template update to improve tracking accuracy.

2. Related Works

Existing target tracking algorithms can be roughly divided into two categories: One is the traditional target tracking method based on correlation filtering [9–13], which uses the response diagram between the template frame and the detection frame after Fourier transform to determine the target of the detection frame. The other is the deep-learning target tracking method based on a convolutional neural network [14–17], which obtains the features of the target by convolutional operation on the images of the template and detection frames, and then obtains the tracking target by similarity matching. This section reviews the related work of researchers in recent years.

2.1. Algorithm Based on Correlation Filter

The tracker based on a discriminant correlation filter (DCF) can effectively use limited data and enhance the training set by using all shifts of local training samples in the learning process. The method based on DCF trains the least-squares regression to predict the target confidence score by using the characteristics of cyclic correlation and fast Fourier transform (FFT) in the learning and detection steps [18]. Mosse [9] was the first pioneering work to propose correlation filter for tracking that uses a random affine set of samples from a single initial frame transformation to construct a minimal output sum of the squares' filter. KCF reduces storage and calculation [6] by several orders of magnitude by diagonalizing the cyclic data matrix with discrete Fourier transform. The periodic assumption of KCF also introduces an unnecessary boundary effect, which seriously reduces the quality of tracking model. SRDCF introduces a spatial regularization component in the learning process in order to reduce the boundary effect, which punishes them according to the spatial position of the correlation filter coefficients [19]. In addition, there are several excellent trackers based on correlation filters, such as STRCF [20] and ECO [8]. They usually divide tracking into two stages: feature extraction and target classification, so end-to-end training is not possible. The objective function in the target classification module in ATOM [13] is based on the mean square error, like the discriminant correlation filtering method, but it is established

on a two-layer fully convolutional neural network. ATOM returns the size of the target with IoU-Net. Although ATOM has achieved effective performance, it is sometimes wrong in size estimation because the predicted joint cross (IoU) may be inaccurate, which leads to tracking failure, especially in a cluttered background.

2.2. Algorithm Based on Siamese Network

In recent years, the visual tracker based on a Siamese network has attracted much attention due to its good balance between tracking performance and efficiency [21]. A Siamese network learns similarity measure functions offline from image pairs, and transforms a tracking task into a template matching task. SiamFC [21] uses a large number of templates and search areas for sample matching in offline training. During online tracking, the template and search regions are correlated in the feature space through forward propagation, and the target position is determined according to the peak position of the correlated response. SiamRPN [12] adds a region proposal network (RPN) to obtain various aspect ratio candidate target frames. It interprets the template branch in a Siamese network as a training parameter, predicts the kernel of a local detection task, and regards the tracking task as a one-time local detection task. SiamMask [22] added a segmented branch based on SiamFC and SiamRPN. The size and shape of the target are obtained, and the tracking results are refined according to the mask of the position of the maximal classification score. One disadvantage of Siamese method is that it ignores the context information around the template, and only extracts the template information from the initial target area.

Due to unrestricted video conditions such as illumination changes and viewpoint changes, the appearance of subsequent targets may be greatly different from that of the initial target. Therefore, the previously proposed Siamese-based tracker degenerates when similar disturbances and object appearance changes occur, which leads to tracking drift and failure. In order to overcome the shortcomings of the Siamese method, DiMP [23] trains a discriminant classifier online and separates the target from the background. This model is derived from a discriminant learning loss by designing a special optimization process, which predicts a strong model in several iterations. The tracker continuously collects positive and negative samples in the tracking process when the target has sufficient confidence prediction, and the classifier template is updated online when 20 frames of target are tracked or a disturbance peak is detected to deal with the appearance change.

So far, many researchers have studied the occlusion problem. However, most of the research is based on the correlation filtering algorithm using handcrafted features, and the effect is not very good. As shown in Figure 1, occlusion may lead to target loss, so it is difficult to achieve accurate tracking through the method based on a Siamese network in actual industrial production. Target redetection algorithms are more used to solve the occlusion problem in tracking. However, the premise is that other cameras must capture unoccluded targets at a time of occlusion, which means that a target requires at least two or more cameras; that is, the number of cameras should at least double. Therefore, it is necessary to propose a cheap method to solve the problem of target occlusion.

Figure 1. Diagram of tracking failure caused by occlusion.

3. Anti-Occlusion UAV Tracking Algorithm by Integrating Attention Mechanism

In order to solve the problem of complex background and occlusion in low-altitude UAV tracking, this paper proposes a single-target tracking algorithm with attention mechanism and anti-occlusion ability. Extracted convolutional features are enhanced to solve

the problem of complex backgrounds by adding a squeeze-and-excitation (SE) module to feature extraction network for feature optimization. Combining target tracking and UAV flight trajectory prediction, the trajectory prediction module is started to predict the position of a UAV when it is occluded. In this paper, the ATOM algorithm is improved. The sequence and exception (SE) module was added to the feature extraction part, and the occlusion-sensing and trajectory-prediction modules were added to the tracking process to realize the robust tracking of low-altitude UAVs.

3.1. Squeeze-and-Excitation (SE) Module

A squeeze-and-excitation (SE) module is an attention method to improve the correlation between feature channels, and enhance target features. By introducing the SeNet attention mechanism, the representation of targets in the channel dimension is enhanced. At the same time, by emphasizing the target and suppressing background information, adjusting the parameters in the network, it shows obvious advantages in image classification. Therefore, the SeNet attention mechanism was added to the ResNet-18 feature extraction network and the final output to enhance the extracted target features in order to solve the complex background problem of low-altitude UAVs.

As shown in Figure 2, squeeze-and-excitation (SE) modules generate different weight coefficients for each channel according to relationship between feature channels, multiplying the previous features and adding them to the original features to achieve the purpose of enhancing features.

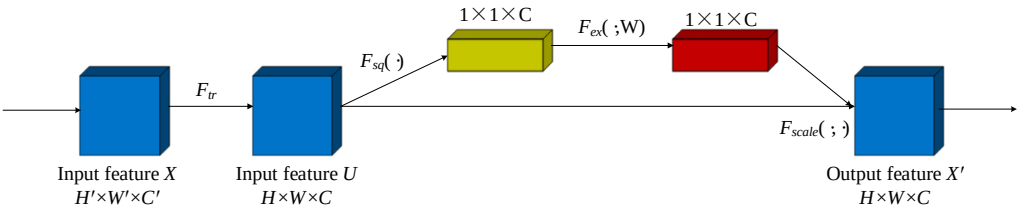

Figure 2. Squeeze-and-excitation (SE) module.

As shown in Figure 2, the process of the SeNet attention mechanism is as follows: First, the extracted feature $X \in \mathbb{R}^{H' \times W' \times C'}$ is mapped to $U \in \mathbb{R}^{H \times W \times C}$ by transforming function Ftr. Then, the global information of each channel is represented by a channel feature description value through global average pooling $Fsq(\cdot)$, and the channel feature description value is adaptively calibrated by $Fex(\cdot, W)$ to render the weight more accurate. Lastly, the enhanced feature $Y \in \mathbb{R}^{H \times W \times C}$ is obtained by multiplying the weight value and the original feature by $Fscale(\cdot, \cdot)$.

Specifically, Ftr is treated as a convolutional operator, $V = [v_1, v_2, \cdots, and v_C]$ represents the set of learned filter kernels, where v_i represents the parameters of the $i - th$ filter. So, the output of X through Ftr is $U = [u_1, u_2, \cdots, u_C]$,

$$u_i = v_i * X = \sum_{t=1}^{C'} v_i^t * x^t \tag{1}$$

where $*$ represents a convolutional operation, and v_i^t is a two-dimensional spatial kernel that represents the channel corresponding to a single channel in X.

Global average pooling $Fsq(\cdot)$: In order to better represent the features of all channels without losing any features, global average pooling is used for the feature information of each channel, and the feature information of the channel is expressed as a value. z_i represents the feature description value of each channel, which is expressed as

$$z_i = F_{sq}(u_i) = \frac{1}{H \times W} \sum_{r=1}^{H} \sum_{c=1}^{W} u_i(r, c) \tag{2}$$

where u_i is a feature in the i-th channel.

Adaptive calibration $Fex(\cdot, W)$: two fully connected layers are used to fully exploit the correlation between channels. First, the number of channels is reduced to C/r through a fully connected layer to reduce the amount of calculation, and the ReLU function is used to activate the output. Then, the number of channels is again restored to C through a fully connected layer, and a sigmoid activation function is adopted to output. This process is expressed as

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \tag{3}$$

where δ is the ReLU activation function, σ is the sigmoid activation function, and $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$,$W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$.

Lastly, enhanced features on the channel are obtained by multiplying the weight coefficient through the fully connected layers with the previous features.

$$y_i = F_{scale}(u_i, s_i) = s_i u_i \tag{4}$$

where y_i is the feature of the $i - th$ channel after weight multiplication, $Y \in [y_1, y_2, \cdots, y'_C]$ is the enhanced feature through the channel.

Inspired by SeNet, combined with the characteristics of ResNet-18 network, a ResNet-18 network combined with SeNet is proposed. On the basis of the original ResNet-18 network, the SeNet layer was added behind each dense block to realize the utilization of the attention mechanism of ResNet-18 network channel. The specific network framework is shown in Figure 3.

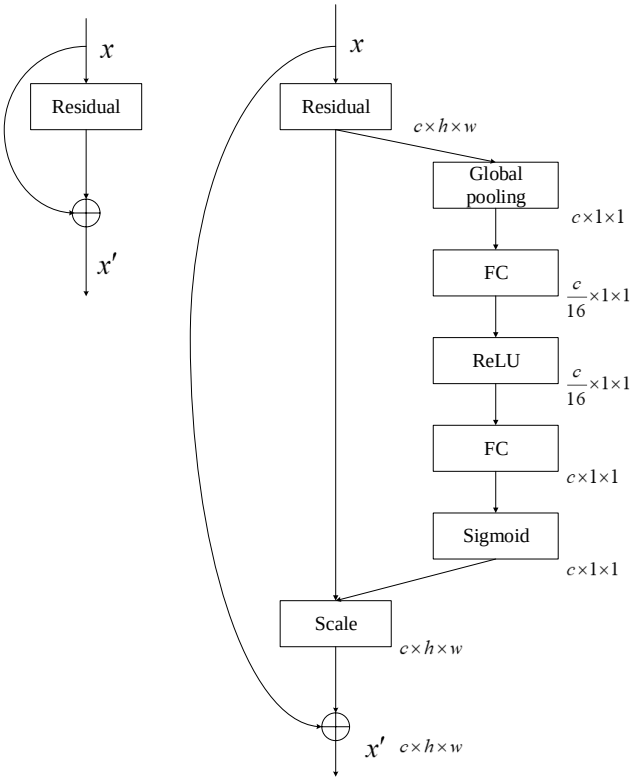

Figure 3. Se-ResNet network by integrating SeNet.

At the same time, after features are extracted from the Se-ResNet network, SeNet is used again to enhance the extracted features to solve the complex background problem in the tracking of low-altitude UAV. The feature extraction network is shown in Figure 4.

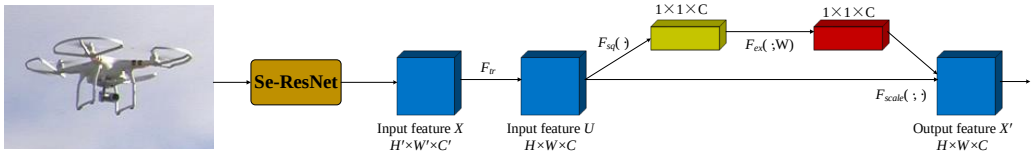

Figure 4. Feature extraction network by integrating attention mechanism.

3.2. Occlusion-Sensing Module

An occlusion-sensing module is proposed to determine whether a target is occluded. The Gaussian response map is obtained by cross-correlation between the feature map from the feature extraction network in the search area and the target frame. Response values within a certain range are found, and the position is recorded as set A. The Euclidean distance between the target and the elements in set A is calculated, and the average value is obtained. If the distance is greater than the threshold set by the algorithm, it is determined as an occlusion.

In this study, the feature response diagram of a feature extraction network was analyzed through the visualization of the training process. When occlusion occurs, the response graph fluctuates and the response peak is not prominent. On the basis of this phenomenon, an occlusion-sensing module is proposed to accurately determine whether the target is occluded, as shown in Figure 5.

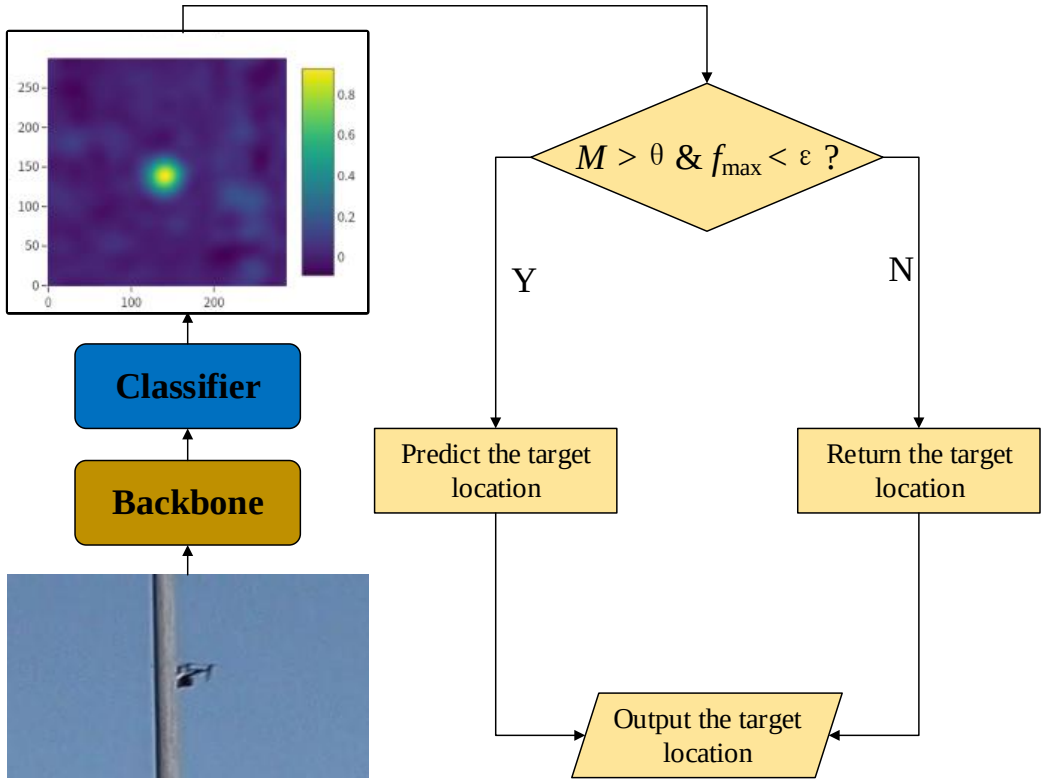

Figure 5. Occlusion-sensing module.

First, points with a certain range of response values in the response diagram are obtained, and the position set is denoted as A.

$$A = \{(i,j)|(\hat{r}(i,j) > \eta_1 mean(\hat{r}))and(\hat{r}(r,j) < \eta_2 max(\hat{r}))\} \quad (5)$$

where $\hat{r}$ is the current frame response graph, and $mean(\hat{r})$ is the average response graph.

Average occlusion distance metric M_O is defined as

$$M_O = \frac{1}{n} \sum_{(i,j) \in A} \sqrt{(i-m)^2 + (j-n)^2} \tag{6}$$

where n represents the number of points contained in set A, and (m, n) represents the location of the peak response.

Figure 6 shows the target response diagram after using the proposed occlusion strategy. Three common response diagrams of a tracking target state are given, namely, no occlusion, partial occlusion, and complete occlusion. The dark points in the figure represent the points in set A. With the increase in the occlusion degree of the target, the response diagram dramatically changed, the number of points in set A increased, the average occlusion distance metric M_O also increased, and multiple peaks appeared in the response diagram. Therefore, average occlusion distance measure M_O could reflect the occlusion state of the target to a certain extent.

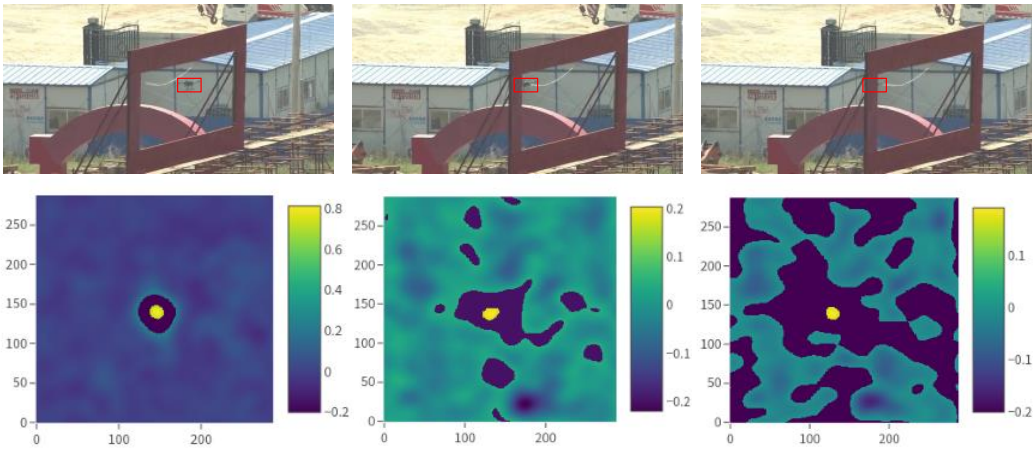

Figure 6. Response diagram under occlusion-sensing module.

In view of the phenomenon described in Figure 6, an occlusion-sensing module is proposed to discriminate the occlusion of UAVs during tracking. The specific process is shown in Figure 5. Feature extraction is performed on the target frame, and a mask operation is performed on the extracted features and the trained template to obtain the Gaussian response diagram. When global mean occlusion distance D_O is greater than the set threshold θ, and the peak f_{max} of the response graph is less than set threshold ε, the UAV is judged to be occluded. The trajectory prediction module is called to predict the next position of the UAV, and the template update is stopped to prevent the template from being occluded.

3.3. UAV Trajectory Prediction Based on LSTM

The traditional Kalman filter algorithm has achieved good results in terms of trajectory prediction and has been applied in engineering. However, the Kalman filter is only applicable to tracking linear moving targets. The single-target tracking problem with different trajectory types is difficult. The measurement value is uncertain, especially when the target is occluded or has disappeared, so it is difficult to effectively predict in this case.

Most target trajectories do not follow the linear principle in common UAV flight videos, which hinders the Kalman filter from predicting trajectories well, while long short-term memory (LSTM) performs better. LSTM is more suitable for solving the prediction problem of a nonlinear motion trajectory because it benefits from its internal mechanism. For example, the Social-LSTM algorithm achieved good trajectory prediction performance. In view of the diversification of target trajectories, a trajectory prediction model is proposed by improving the LSTM algorithm. The central coordinates of the historical frame before

occlusion are used as the input of the trajectory prediction model, trajectory samples are generated by LSTM, and the next prediction position of the target is obtained, which solves the problem of tracking failure when the UAV is occluded.

The space coordinate of UAV at the time t is (x_t, y_t), in which the time $t = 1$ to $t = t_{obs}$ is observable, and the corresponding observable trajectory is represented as $(x_1, y_1), (x_2, y_2), \cdots, (x_{obs}, y_{obs})$. Time t=t_{pred} is the prediction time, and the corresponding prediction coordinates is represented as $(\hat{x}_{pred}, \hat{y}_{pred})$.

The trajectory prediction network based on LSTM proposed in this paper is shown in Figure 7. With the historical flight trajectory of UAV as the input, the predicted flight trajectory is output after a LSTM encoder and decoder.

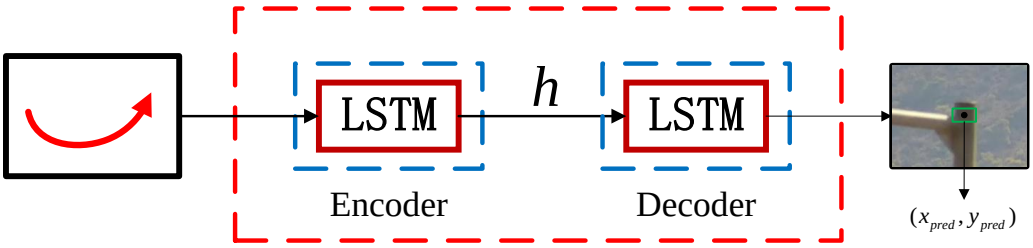

Figure 7. Trajectory prediction network based on LSTM.

Let h^t represent the hidden state of LSTM at time t, which is used to predict the distribution of target position $(\hat{x}_{t+1}, \hat{y}_{t+1})$ at time $t + 1$. Assuming that it obeys binary Gaussian distribution, mean μ_{t+1}, standard deviation σ_{t+1} and correlation coefficient ρ_{t+1} are predicted by weight matrix W_p. Then, prediction coordinate $(\hat{x}_t, \hat{y}_t)$ at time t is:

$$(\hat{x}_t, \hat{y}_t) \sim (\mu_t, \sigma_t, \rho_t) \tag{7}$$

The parameters of the model are learnt by minimizing the negative logarithmic likelihood function:

$$[\mu_t, \sigma_t, \rho_t] = W_p h^{t-1} \tag{8}$$

$$L(W_e, W_l, W_p) = -\sum_{t=T_{obs+1}}^{T_{pred}} \log(P(x_t, y_t|\sigma_t, \mu_t, \rho_t)) \tag{9}$$

The model is trained by minimizing the loss for all trajectories in the training dataset, where W_l is the network weight of the LSTM, and W_e is the weight of the position coordinates. Because this article only predicts the trajectory of the UAV, there is no relationship with other trajectories, and there is no need to calculate the weight associated with other trajectories, so the weight W_e of the position coordinate iwass set to 1.

3.4. Comprehensive Scheme and Algorithm Implementation

Anti-occlusion target tracking for UAVs integrating the SeNet attention mechanism is proposed considering the SeNet attention module, occlusion-sensing module, and flight trajectory prediction above, as shown in Algorithm 1. Global and local information is fused for feature enhancement by using the SeNet attention mechanism. The occlusion-sensing module is used to determine whether the target is occluded. When the target is occluded, the LSTM algorithm is used to predict the target position. The whole process of the algorithm is shown in Figure 8.

Algorithm 1 Proposed UAV tracking algorithm.

Input: Target position pos and the size of bounding box $rect$ in the first frame.
Output: Target position pos_i and the size of bounding box $rect_i$ in the $i-th$ frame.

Initialize N_{image}, $pooling$, t, ϵ.
for $i = 2$ to N_{image} **do**
  Extract the area of $pooling * rect$ size as search area with pos_{i-1} coordinates in the $i-th$ frame.
  Extract features in search area by the backbone.
  Generate response graph using classified regression filter.
  Calculate A and M_O using Equations (5) and (6).
  if $M_O > \theta$ *and* $f_{max} < \varepsilon$ **then**
    Call LSTM trajectory prediction algorithm, enter $[pos_{i-t}, ..., pos_{i-1}]$, and output pos_i.
  else
    Output classification regression filter response graph corresponding position pos_i.
  end if
  Extract multiple bounding boxes of different scales with pos_i as the coordinate origin, and calculate the IoU scores. The bounding box with the highest score corresponds to the pos_i and $rect_i$ of the target in the $i-th$ frame.
end for

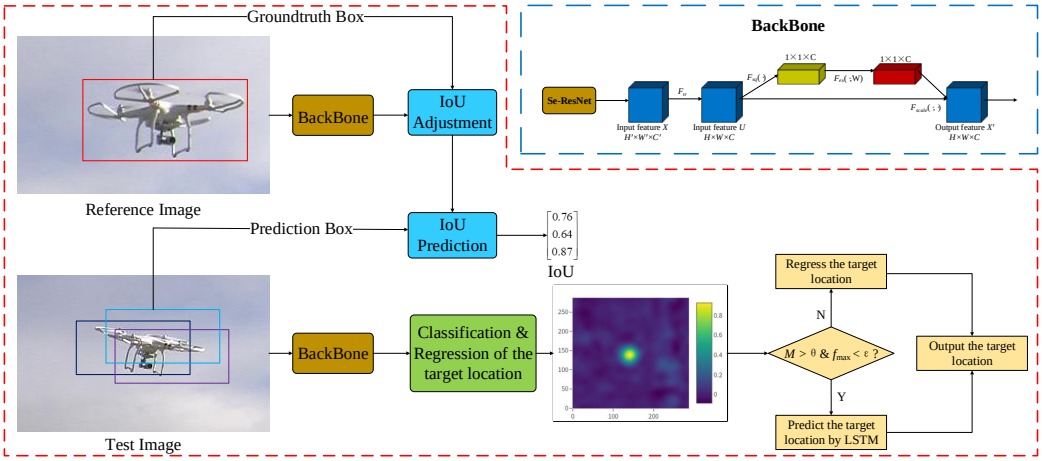

Figure 8. Comprehensive scheme of the proposed UAV tracking algorithm.

4. Experimental Results and Analysis

In this paper, some UAV datasets in Drone-vs.-Birds [24] and LaSOT [25] were integrated to form UAV datasets. Experimental verification was carried out on the OTB-100, GOT-10k, and integrated UAV datasets to verify the effect of the improved algorithm proposed in this paper, and the tracking-success and precision plots were used for evaluation.

4.1. Experimental Environment and Parameters Setting

The algorithm was implemented in Python 3.7 with the PyTorch framework. The experimental computer operating system was Ubuntu 180.4 64-bit, CPU InterCore i7-9700k, the main frequency was 3.60 GHz, with 16 GB memory, NVIDIA GeForce RTX2080Ti, and 11 GB memory. In the training process, some LaSOT and GOT-10k-train dataset are used as the training set, and the part of GOT-10k-train dataset that does not participate in the training is used as the verification set. The pretraining parameters on ImageNet are used in the backbone. By training the network, the common features in the visual tracking process are learned for the following tracking. In the tracking process, the occlusion threshold is set to $\theta = 12$, $\varepsilon = 0.1$.

4.2. Comparison and Analysis of Experimental Results

4.2.1. Experiment on OTB-100 Dataset

The OTB-100 dataset contains 100 different video sequences. The coordinates of the target and the size of the bounding box in the sequence are manually labeled, and are relatively accurate. The dataset contains 25% gray images, which pose a challenge to the algorithm on the basis of color feature tracking.

The proposed algorithm was tested on OTB-100 and compared with four advanced trackers, namely, the Siamfc, Dimp, Prdimp, and ATOM algorithms. Figure 9 shows the precision and success plots of the five algorithms on the OTB-100 dataset. The precision and success rate of the proposed algorithm were improved compared with the second algorithm after adding the SeNet attention mechanism and anti-occlusion module.

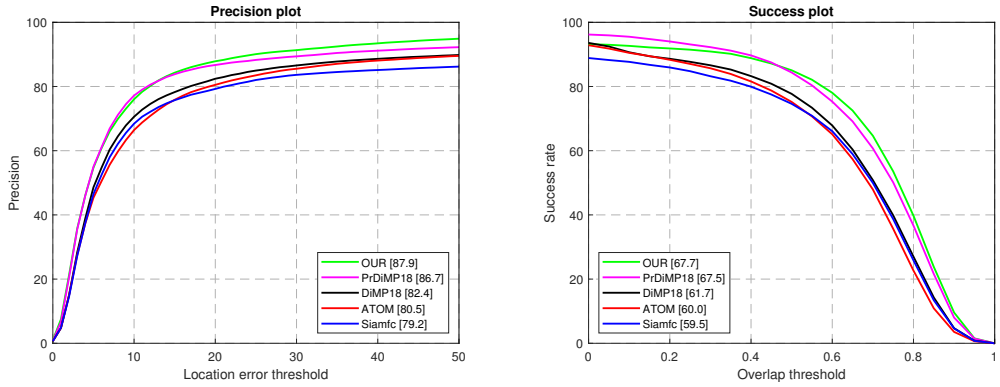

Figure 9. Precision and success plots on OTB-100 dataset.

The tracking results of some video sequences of the OTB-100 dataset are shown in Figure 10. The ATOM and Siamfc algorithms lost the target if the occlusion time was too long. However, the algorithm proposed in this paper could effectively resist occlusion with the use of the anti-occlusion module. Furthermore, for short-term occluded targets, the algorithm proposed in this paper tracked the target position more accurately than other baseline algorithms did because of the LSTM prediction module.

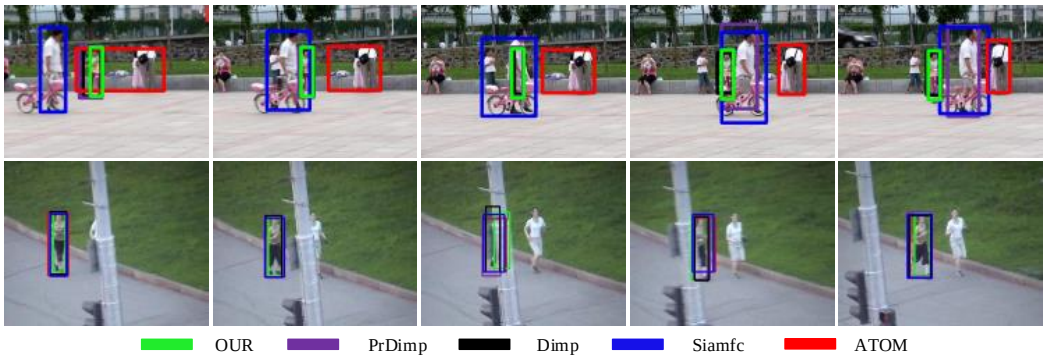

Figure 10. Tracking results of video sequences of the OTB-100 dataset.

4.2.2. Experiment on GOT-10k Dataset

The GOT-10k dataset contains video sequences of more than 10,000 moving targets in the real world, in which more than 1.5 million targets are manually marked in location and bounding box. The GOT-10k test set contains 84 target categories and 32 moving target categories, without overlap between the training set and test set. Therefore, GOT-10k-val for testing is not affected by GOT-10k-train for training.

The proposed algorithm was also compared with the Siamfc, Dimp, Prdimp and ATOM algorithms on the GOT-10k dataset. Figure 11 shows the precision and success plots

of the five algorithms on the GOT-10k dataset. After adding the SeNet attention mechanism and anti-occlusion module, the accuracy and success rate of the proposed algorithm were 59.9% and 73.1%, respectively, which were 8.3% and 3.7% higher than those of the second algorithm.

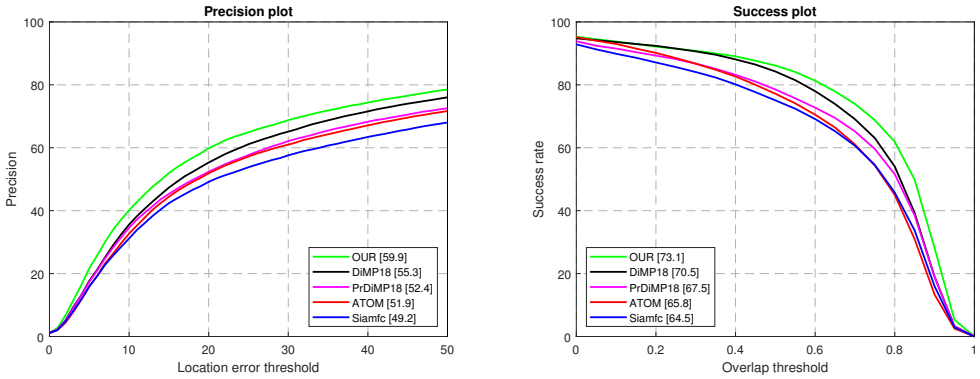

Figure 11. Precision and success plots on GOT-10k dataset.

The visualization of part of the video sequence tracking results of the GOT-10k-val dataset is shown in Figure 12. The ATOM and Siamfc algorithms were not accurate in predicting the target scale in a complex background. With the use of the SeNet attention mechanism, the algorithm proposed in this paper was more accurate for the scale regression of the target than other baseline algorithms were.

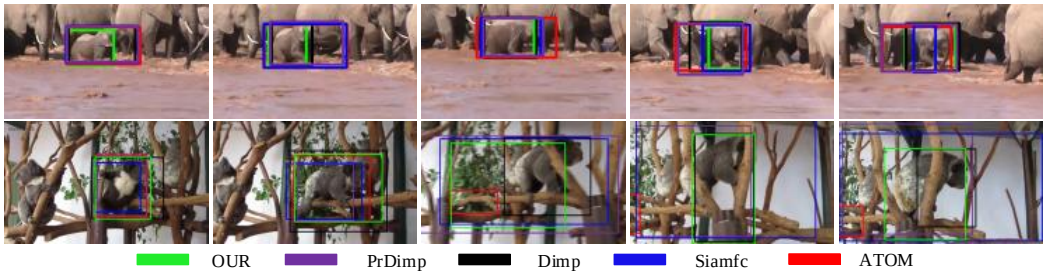

Figure 12. Tracking results of video sequences of the GOT-10k dataset.

4.2.3. Experiment on Integrated UAV Dataset

Drone-vs.-Birds is a target detection dataset used to distinguish between UAVs and birds with video sequences of UAVs and birds. This study uses its UAV video sequence and the UAV video sequence of the LaSOT dataset to form a dataset for UAV tracking to verify the proposed algorithm. UAV video sequences in the Drone-vs.-Birds and LaSOT datasets were combined into a dataset for UAV tracking to verify the algorithm proposed in this study.

Figure 13 shows the precision and success plots of the Siamfc, Dimp, Prdimp, ATOM, and proposed algorithms on the integrated UAV dataset. The accuracy and success rate of the proposed algorithm were 79% and 50.5%, which are 10.6% and 4.9% higher than those of the second algorithm.

The visualization of the partial tracking process is shown in Figure 14. When occlusion occurred, the Siamfc and ATOM algorithms may have lost the target and failed in tracking. With the use of SeNet attention mechanism and anti-occlusion module, the algorithm proposed in this paper could achieve better tracking results.

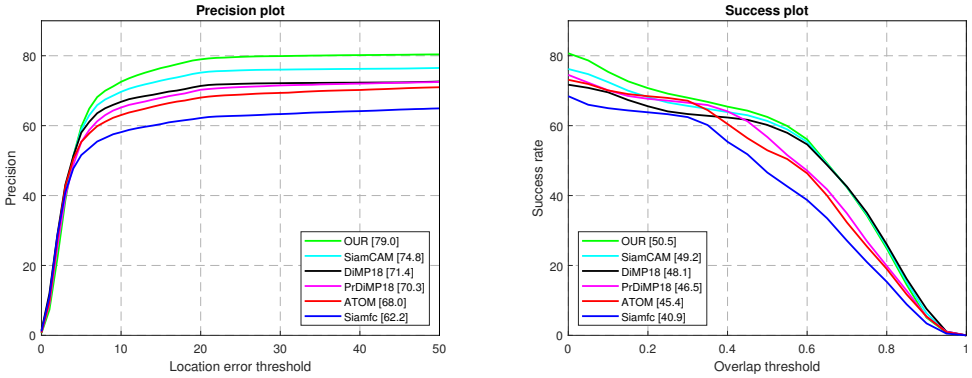

Figure 13. Precision and success plots on integrated UAV dataset.

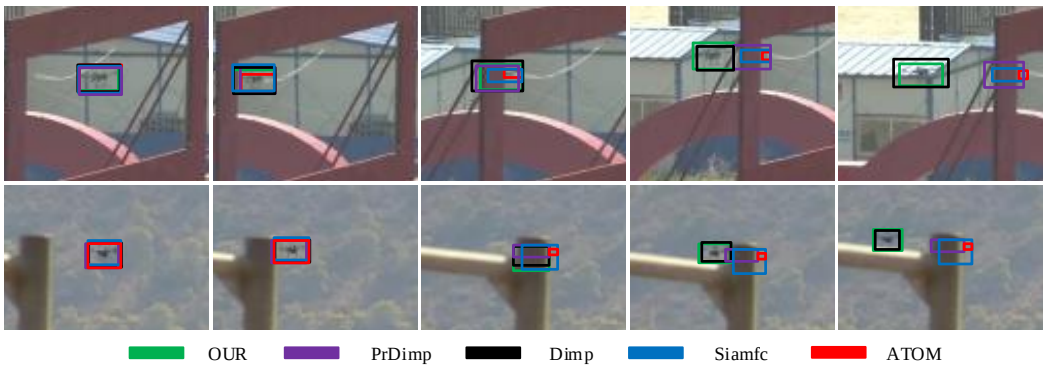

Figure 14. Tracking results under occlusion of the integrated UAV dataset.

5. Conclusions

Aiming at the problems of complex background and occlusion of UAVs in low-altitude airspace during flight, an anti-occlusion UAV tracking algorithm with an integrated attention mechanism was proposed. In this algorithm, the SeNet attention mechanism is introduced to fuse global and local information for feature enhancement to solve the problem of complex backgrounds. The occlusion-sensing module was designed to determine whether the target is occluded, and if the target is occluded, the LSTM algorithm is used to predict the target position to solve the occlusion problem. By validating on three different datasets, the method proposed in this paper achieved good results and tracked UAVs well. However, with the addition of SeNet attention mechanism and anti-occlusion module, the algorithm parameters increased and the amount of calculation increased, resulting in a decrease in the running speed of the algorithm. The running speed on the GPU 2080ti server was 49 fps/s, which basically achieves real-time tracking. Further improving the tracking speed and performance of the algorithm without reducing its accuracy is future research work.

Author Contributions: Conceptualization, C.W. and J.W.; methodology, Z.S.; investigation, Q.G.; resources, C.W. and T.W.; writing—original draft preparation, Z.S.; writing—review and editing, C.W. and L.M.; visualization, E.W.; supervision, C.W. and Q.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China with grant Nos. 61703287 and 62173237, Scientific Research Program of Liaoning Provincial Education Department of China with grant Nos. LJKZ0218 and JYT2020045, Young and middle-aged Science and Technology Innovation Talents Project of Shenyang of China with grant No. RC210401 and Liaoning Provincial Key R&D Program of China with grant No. 2020JH2/10100045.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the reviewers and editors for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tapsall, B.T. Using crowd sourcing to combat potentially illegal or dangerous UAV operations. In Proceedings of the Unmanned/Unattended Sensors and Sensor Networks XII. SPIE, Edinburgh, UK, 27 September 2016; Volume 9986, pp. 23–28.
2. Jin, H.; Wu, Y.; Xu, G.; Wu, Z. Research on an Urban Low-Altitude Target Detection Method Based on Image Classification. *Electronics* **2022**, *11*, 657. [CrossRef]
3. Liu, C.; Xu, S.; Zhang, B. Aerial Small Object Tracking with Transformers. In Proceedings of the 2021 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 15–17 October 2021; pp. 954–959.
4. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ATOM: Accurate Tracking by Overlap Maximization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4655–4664.
5. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
6. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
7. Wu, Y.; Yang, M. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef]
8. Huang, L.; Zhao, X.; Huang, K. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1562–1577. [CrossRef]
9. Yin, H.P.; Chen, B.; Chai, Y.; Liu, Z.D. Vision-based object detection and tracking: A review. *Acta Autom. Sin.* **2016**, *42*, 1466–1489.
10. Li, Y.; Zhu, J. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In Proceedings of the 2014 European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 254–265.
11. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 583–596. [CrossRef] [PubMed]
12. Danelljan, M.; Robinson, A.; Shahbaz Khan, F.; Felsberg, M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 472–488.
13. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6931–6939.
14. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
15. Held, D.; Thrun, S.; Savarese, S. Learning to Track at 100 FPS with Deep Regression Networks. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 749–765.
16. Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H. End-to-end representation learning for correlation filter based tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2805–2813.
17. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
18. Gladh, S.; Danelljan, M.; Khan, F.S.; Felsberg, M. Deep motion features for visual tracking. In Proceedings of the 2016 International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 1243–1248.
19. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4310–4318.
20. Li, F.; Tian, C.; Zuo, W.; Zhang, L.; Yang, M.H. Learning spatial-temporal regularized correlation filters for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4904–4913.
21. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 850–865.
22. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338.
23. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning Discriminative Model Prediction for Trackingh. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 6181–6190.

24. Coluccia, A.; Fascista, A.; Schumann, A.; Sommer, L.; Dimou, A.; Zarpalas, D.; Akyon, F.C.; Eryuksel, O.; Ozfuttu, K.A.; Altinuc, S.O.; et al. Drone-vs-Bird Detection Challenge at IEEE AVSS2021. In Proceedings of the 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Washington, DC, USA, 16–19 November 2021; pp. 1–8.
25. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5374–5383.