remote sensing

MDPI

Article

Learning Spatio-Temporal Attention Based Siamese Network for Tracking UAVs in the Wild

Junjie Chen [1,†], Bo Huang [1,†], Jianan Li [1], Ying Wang [1], Moxuan Ren [1] and Tingfa Xu [1,2,*]

[1] Image Engineering & Video Technology Lab, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China; 3120190516@bit.edu.cn (J.C.); 3120185333@bit.edu.cn (B.H.); lijianan@bit.edu.cn (J.L.); 3120215325@bit.edu.cn (Y.W.); renmoxuanx@bit.edu.cn (M.R.)

[2] Big Data and Artificial Intelligence Laboratory, Beijing Institute of Technology Chongqing Innovation Center (BITCQIC), Chongqing 401135, China

* Correspondence: ciom_xtf1@bit.edu.cn

† These authors contributed equally to this work.

Abstract: The popularity of unmanned aerial vehicles (UAVs) has made anti-UAV technology increasingly urgent. Object tracking, especially in thermal infrared videos, offers a promising solution to counter UAV intrusion. However, troublesome issues such as fast motion and tiny size make tracking infrared drone targets difficult and challenging. This work proposes a simple and effective spatio-temporal attention based Siamese method called SiamSTA, which performs reliable local searching and wide-range re-detection alternatively for robustly tracking drones in the wild. Concretely, SiamSTA builds a two-stage re-detection network to predict the target state using the template of first frame and the prediction results of previous frames. To tackle the challenge of small-scale UAV targets for long-range acquisition, SiamSTA imposes spatial and temporal constraints on generating candidate proposals within local neighborhoods to eliminate interference from background distractors. Complementarily, in case of target lost from local regions due to fast movement, a third stage re-detection module is introduced, which exploits valuable motion cues through a correlation filter based on change detection to re-capture targets from a global view. Finally, a state-aware switching mechanism is adopted to adaptively integrate local searching and global re-detection and take their complementary strengths for robust tracking. Extensive experiments on three anti-UAV datasets nicely demonstrate SiamSTA's advantage over other competitors. Notably, SiamSTA is the foundation of the 1st-place winning entry in the 2nd Anti-UAV Challenge.

Keywords: anti-UAV; single object tracking; thermal infrared videos; Siamese network; correlation filter; re-detect; motion feature

check for updates

Citation: Chen, J.; Huang, B.; Li, J.; Wang, Y.; Ren, M.; Xu, T. Learning Spatio-Temporal Attention Based Siamese Network for Tracking UAVs in the Wild. *Remote Sens.* **2022**, *14*, 1797. https://doi.org/10.3390/rs14081797

Academic Editor: Francesco Nex

Received: 24 January 2022
Accepted: 29 March 2022
Published: 8 April 2022

1. Introduction

In recent years, the rapid development of unmanned aerial vehicles (UAVs) has promoted a large number of applications, such as aerial photography [1–3], video surveillance [4], and biological monitoring [5]. On the contrary, the potential abuses of this technology could lead to significant negative impacts on society. Thus, anti-UAV techniques are of great importance and in urgent need of practical research, among which vision-based approaches are more widely adopted due to their higher efficiency, lower power consumption, and easier deployment.

Given a UAV target specified by a bounding box in the first frame, visual object tracking aims to determine the exact state of the target sequentially in a video, which servers as a fundamental step for anti-UAV task. As is obvious, Thermal Infrared (TIR) tracking technique is better suited to the low-light scenarios, thus catering to all-weather requirements. However, tracking UAVs in TIR video, compared to tracking objects in visible video, further introduces significant challenges, e.g., small object, fast motion, thermal clutter background. How to tackle these problems remain challenging and ill-solved.

Siamese-based algorithms play a dominant role in the field of visual object tracking. SiamFC [6] first treats the tracking task as a similarity matching problem between target template and the search region. Later on, numerous improvements have been done in terms of adding auxiliary branch [7], digging deeper feature [8], improving embedding strategy [9], augmenting online update mechanism [10–12], etc. While considering the tracking object is UAV, which contains a wide range of fast motion and out-of-view situations, unlike the improvements mentioned above, the global re-detection mechanism and trajectory modeling could play critical roles in accurate tracking.

SiamRCNN [13] introduces global re-detection mechanism into Siamese networks and achieves outstanding tracking performance, which is thus used as our baseline algorithm. But on the other hand, as the UAV targets themselves severely lack semantic features, continuous global detection mechanism could be more likely to induce tracking drift, especially when the UAV targets are drowned in thermal clutter background. At this time, it seems more appropriate to detect targets in the local neighborhoods. Obviously, local detection and global re-detection are the two opposite strategies that can cope with different challenging situations during tracking. For this reason, this work elaborately designs a framework adaptively switching these two strategies, achieving robust tracking through performing reliable designed local tracking and wide-range re-detection alternatively.

This paper proposes a simple yet effective spatio-temporal attention-based Siamese network, called SiamSTA, to track UAVs in TIR videos robustly. SiamSTA follows the typical Siamese framework that consists a template branch and a detection branch. The template branch extracts features for the template target specified in the first frame, while the detection branch takes the search image as input and selects target candidates from redundant Region Proposal Network (RPN) proposals. To tackle the key challenges, i.e., small scale and fast movement, commonly faced in anti-UAV tracking scenarios, SiamSTA integrates both a reliable local tracking and a wide-range global re-detection mechanism, and takes their complementary advantages in an alternative-performing fashion.

Specifically, to better perceive small targets that easily be distracted by background clusters, the local tracking strategy incorporates spatial and temporal constraints to limit the position and aspect ratio of generated candidate proposals in a local neighborhood, so as to suppress background distractors and locate the target accurately. Meanwhile, in case the target runs out of the local region due to rapid movement, a three-stage global re-detection mechanism is designed to redetect the target: (i) provides re-detections using the first-frame template, (ii) implements re-detections of high-confidence predictions from previous frames, and (iii) adopts correlation filter based on change detection, short for CDCF, to exploit beneficial motion features to better locate fast-moving target in a wide range. Finally, a switching policy is adopted to apply local tracking and global re-detection adaptively depending on varying target states to make optimal predictions, hence achieving robust tracking.

To verify the performance of SiamSTA, comprehensive experiments are conducted on three challenging UAV infrared tracking datasets, i.e., Anti-UAV2020 test-dev [14], Anti-UAV2021 test-dev [15], Anti-UAV [16]. Detailed experimental comparisons show that SiamSTA has advantages over its competing counterparts in addressing the key challenges faced by anti-UAV tracking, including but not limited to small scale and fast movements. In addition, SiamSTA serves as the foundation of the 1st-place winning entry in the 2nd Anti-UAV Challenge, further evidencing its robustness in real-world scenarios.

To sum up, this paper makes the following contributions:

- This paper proposes a novel Siamese based tracker that integrates local tracking and global re-detection mechanisms in a unified framework and perform them adaptively depending on varying target states.
- This paper designs a spatio-temporal attention based local tracking strategy to eliminate background clusters and better perceive small targets.
- A three-stage global re-detection strategy to recapture targets in a wide range is proposed.

- Our method establishes state-of-the-art performance in Anti-UAV2020 test-dev [14], Anti-UAV2021 test-dev [15] and Anti-UAV [16] datasets.

2. Related Works

This section first reviews the development of two mainstream tracking frameworks, i.e., correlation filters and Siamese networks. Then, some representative tracking algorithms in thermal infrared videos are introduced.

2.1. Correlation Filter

Correlation filter (CF) based tracking methods have been widely applied to visual tracking due to the high computational efficiency. MOSSE [17] tracker is the first to validate the feasibility of the correlation filter in tracking. After MOSSE, CF trackers obtain a wide attention. Refs. [18,19] introduce circulant matrix to produce enough samples for training, while maintaining a fast speed. However, the resulting periodic repetitions at boundary positions limit the discriminative ability of the tracker. To mitigate this issue, Danelljan et al. [20] adopt a spatial regularization term which allows the tracker focus on the target center, Galoogahi et al. [21] apply a binary matrix to crop real samples for training, both of which are promising in lifting the performance of trackers. Besides addressing boundary effect, some CF trackers introduce scale estimation [22–24] to improve the tracking performance. In addition, A powerful feature extraction such as HOG [18,19], CN [25] and deep feature [26] will enhance the feature representation ability and increase the tracking accuracy. However, traditional CF trackers mostly apply nearby search, which is difficult to capture the target again when the target is out of view for a while and then re-enters the field of view, which limits the application of CF trackers in anti-UAV mission.

2.2. Siamese Network

Recently, the Siamese network based trackers have gained a lot of attention for their great success in multiple video object tracking benchmarks and competitions. Bertinetto et al. [6] propose the initial SiamFC tracker, which formulates visual tracking as a cross-correlation problem and expects to learn a similarity evaluation map based fully-convolutional network in an end-to-end manner. Li et al. [27] significantly enhance the tracking performance of SiamFC by introducing a Region Proposal Network (RPN), which allows estimation of target locations, sizes, and ratios by enumerating multiple anchors. However, these trackers implement nearby search, which is difficult to recapture target after it lost. Recently, Voigtlaender et al. [13] unleash the full power of global searching by a two-stage Siamese re-detection architecture, which makes full use of both the first-frame template and previous-frame predictions for the optimal decision. Ref. [13] not only solves the problem of update, but also improves the probability of re-detection after target lost. However, global searching also introduces too much distractors which hurts the performance of tracking small target in complex background. To this end, this paper proposes a spatio-temporal attention based Siamese network to enhance the tracking robust of global re-detection.

2.3. TIR Tracking

Recently, more attentions [28–30] have been paid to TIR tracking for the rapidly development of infrared sensors in resolution and quality. Due to the poor semantic information in TIR images, how to extract effective features is crucial to distinguish targets from background. Refs. [31–33] compute motion features by thresholding the absolute difference between the current and the previous frame in pixel-wise as an extra feature channel, which is beneficial for identifying moving objects. Yu et al. [34] propose structural learning on dense samples around the object. Their tracker uses edge and HOG features which is suitable for UAV tracking. With the development of deep learning, Convolutional Neural Networks (CNN) have shown competitive performance compared to handcrafted feature. However, due to the limited semantic information in TIR images, traditional

RGB backbones performs poorly, and a number of works [35,36] start to design networks specialized for TIR images. However, the limited data in the infrared dataset and the large amount of data required for training the deep network are not friendly to anti-UAV tasks. Wu et al. [36] exploit the information in the initial frame to train a feature extraction network for correlation filter. However, in long-term tracking, the target and background in the initial frame and subsequent frames differ significantly, which is not general to long-term anti-UAV tasks. Based on our previous work [37], our method first extracts motion features using a Gaussian mixed model to select candidate regions where small targets are likely to exist, and then combines the Histograms of Oriented Gradients (HOG) features of the candidate regions to train a Correlation Filter based tracker to assist the Siamese network in making joint decisions, thus better perceiving small targets in TIR images. Finally, a novel distractor-aware regularization term is further proposed to learn the distractor information in the background, thus leading to better tracking robust against thermal clutter background.

3. Method

This section first briefly reviews the baseline tracker SiamRCNN [13]. Then the design of the proposed SiamSTA is described, which consists of spatio-temporal constraints, global motion estimation and change detection based CFs. Next, the online tracking and updating strategy integrating both local search and global detection is further presented.

3.1. Revisiting SiamRCNN

SiamRCNN [13] is a two-stage Siamese tracker with elaborate re-detection mechanism. Its network architecture is sequentially composed of three modules: (1) A backbone feature extraction module containing a template branch for extracting ground-truth features in the target area, and a test branch for preparing possible RPN proposals in the search area; (2) A re-detection head module which performs a two-stage re-detection to learns a similarity evaluation using the initial template and previous predictions; (3) An online dynamic programming module that implicitly tracks both the target and potential similar-looking distractors based on spatio-temporal cues. In the vital third module, SiamRCNN preserves plenty of discontinuous trajectories for making the most comprehensive decisions. Suppose one tracking trajectory consists of N non-overlapping sub-trajectories, $A = (a_1, a_2, \dots, a_N)$, each sub-trajectory $a_i, \forall i \in \{1, 2, \dots, N-1\}$ satisfies $end(a_i) < start(a_{i+1})$, where $start$ and end denote the start and end times of a sub-trajectory, respectively. The overall measuring score of such trajectory is computed by,

$$score(A) = \sum_{i=1}^{N} sim\_eva(a_i) + \sum_{i=1}^{N-1} w_l loc\_eva(a_i, a_{i+1}), \tag{1}$$

where the similarity evaluation sim_eva and location consistency evaluation loc_eva are defined as following,

$$sim\_eva(a_i) = \sum_{t=start(a_i)}^{end(a_i)} [w_r sim(a_{i,t}, gt) + (1 - w_r) sim(a_{i,t}, a_{i,start})], \tag{2}$$

$$loc\_eva(a_i, a_{i+1}) = -|end\_box(a_i) - start\_box(a_{i+1})|_1, \tag{3}$$

here w_l and w_r are the complementary ratios. $a_{i,t}$ denotes the detection of sub-trajectory a_i at time t, and $a_{i,start}$ means the first detection of a_i. $sim(a_{i,t}, gt)$ and $sim(a_{i,t}, a_{i,start})$ return the re-detection confidence of $a_{i,t}$ using the first-frame ground truth reference and the initial detection of the current sub-trajectory, respectively. As presented in Equation (3), the location consistency evaluation between two adjacent sub-trajectory is computed using the negative L_1 norm of the difference between the last bounding box of a_i and the first bounding box of a_{i+1}.

SiamRCNN backs up a lot of trajectories to ensure the success rate of re-detection. However, due to the lack of semantic target features and the presence of complex thermal background in TIR video, these trajectories introduce a large number of similar-looking distractors, thus causing too much disruptions during tracking and eventually leading to tracking drift. To address such issue, finer exploitation of spatio-temporal prior knowledge is a feasible solution.

3.2. SiamSTA Framework

Inspired by SiamRCNN, SiamSTA is built based on a three-stage re-detection mechanism that first retains template information in the initial frame, then integrates predictive information from historical frames, and finally lifts discriminative capability to perceive tiny objects with a change detection based CF, as shown in Figure 1. To deal with background distractors, several practical guidelines using spatio-temporal attention are introduced to regulate candidate proposals. SiamSTA further incorporates a collaborative strategy that combines local search and global detection to facilitate online tracking.

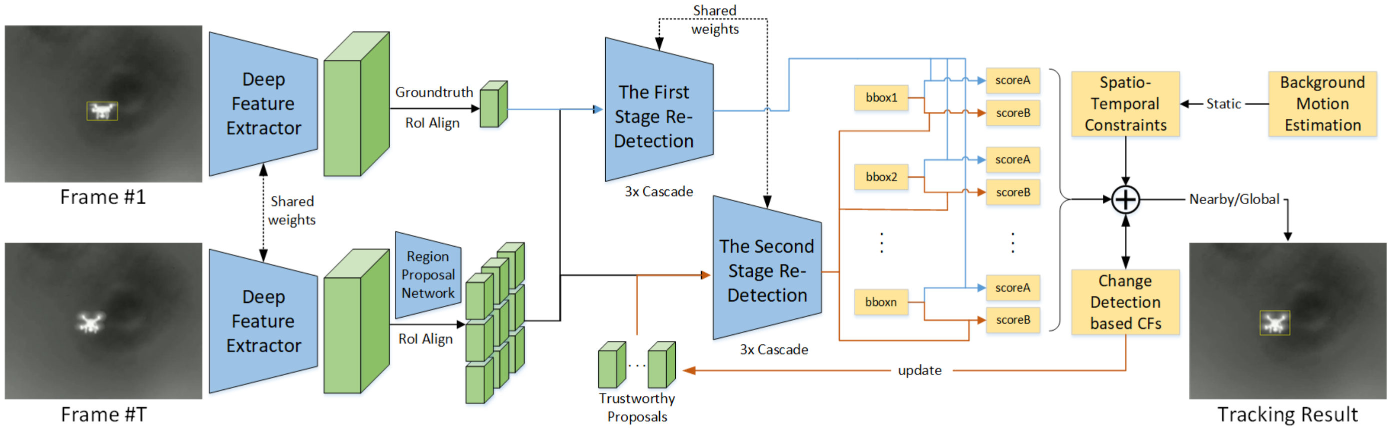

Figure 1. Overall architecture of SiamSTA. It consists of a Siamese backbone that extracts deep features from the template and the search image, followed by a three-stage re-detector that first re-detects the first-frame template, then re-locates historical predictions from previous frames, and finally fixes potential tracking failures using a change detection based CF. The symbol $\oplus$ indicates an ensemble classifier that conditionally switches between local track and global detection to make optimal decisions upon predictions from the three-stage re-detector.

3.2.1. Spatio-Temporal Constraints for Local Tracking

UAV targets in practical TIR tracking are typically very small and without salient texture or fixed shapes, making them extremely hard to be distinguished. To alleviate this, a novel spatio-temporal constraint is introduced. From the spatial perspective, considering the drastic position changes of the targets are unlikely to occur in two adjacent frames captured by a long-range static camera, reliable tracking results can be obtained by searching for targets in local neighborhoods rather than detecting it globally. From the temporal perspective, SiamSTA introduces a memory bank to store valuable historical states of the targets, i.e., target size and aspect ratio, learned from all previous frames to better distinguish potential distractors.

Concretely, SiamSTA records the historical minimum and maximum size and aspect ratio of the target appeared in all previous frames, denoted as (S_{min}, S_{max}) and (R_{min}, R_{max}), respectively, to indicate its range of potential scale variation. Initialize $S_{min} = S_{max} = S$, $R_{min} = R_{max} = R$ with the size S and aspect ratio R of the ground-truth target bounding box specified in the first frame. For an arbitrary frame c, we specify a local neighborhood around the previous target center as the search region where the target is most likely to appear. Only if a high-confidence proposal has been found within the specified search region, whose size S_c and aspect ratio R_c meets the constrain below $S_c \in [0.8 * S_{min}, 1.2 * S_{max}]$,

$R_c \in [0.8 * R_{min}, 1.2 * R_{max}]$, we regard the detection result to be reliable and the trajectory to be continuous. Then SiamSTA updates the stored target state as following:

$$S_{min} = min(S_{min}, S_c), S_{max} = max(S_{max}, S_c), \tag{4}$$

$$R_{min} = min(R_{min}, R_c), R_{max} = max(R_{max}, R_c). \tag{5}$$

The above process lasts until the end of a trajectory. Define the trusted trajectory as $C = (c_1, c_2, \ldots, c_L)$, and compute the evaluation score of a candidate proposal cc as,

$$score(cc) = w_r sim(cc, gt) + (1 - w_r)\frac{1}{L}\sum_{i=1}^{L} sim(cc, c_{i,start}) + w_l iou(cc, c_{L,end}), \tag{6}$$

where $iou(cc, c_{L,end})$ is the intersection over union (IoU) of $bbox(cc)$ and $bbox(c_{L,end})$. Thanks to the spatio-temporal constraints, the number of remaining candidate proposals can be very small, or even unique, which greatly alleviates the interference of distractors.

However, if the target is temporarily lost, the local search strategy may cause the tracker to fail completely. To mitigate the effect of target loss, especially severe occlusion or out-of-view, global re-detection techniques associated with a mutual compensation mechanism that conditionally switches between local tracking and global search is essential, as detailed below.

3.2.2. Global Motion Estimation

Targets in anti-UAV tracking are typically very small with little semantic information, which easily leads to early tracking failures. Fortunately, background scenes in such tracking scenario commonly remain fixed throughout an entire sequence, which provides feasibility to employ motion features to re-capture lost targets.

Motivated by this, this section establishes a global motion estimation model to reveal dynamic change of background scenes. To be specific, SiamSTA extracts the ShiTomasi [38] key points from background regions and track these points to estimate the motion of background scenes. Let $I(x, y)$ denotes the intensity value of pixel (x, y) on input image I. Key points should have a significant gradient change in gray values, such as corner points. Let $[u, v]$ be the local displacement, and the gradient change vector in the local neighborhood can be calculated as,

$$E(u, v) = \sum_{x,y} \tau(x, y)[I(x + u, y + v) - I(x, y)]^2, \tag{7}$$

where $\tau(x, y)$ is a Gaussian window function. Equation (7) can be further simplified as,

$$E(u, v) \cong [u, v]M\begin{bmatrix} u \\ v \end{bmatrix}, \tag{8}$$

where M is a 2×2 matrix:

$$M = \sum_{x,y} w(x, y)\begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \tag{9}$$

where I_x and I_y represent the derivatives of image I in the horizontal and vertical direction, respectively. We can obtain two eigenvalues λ_1, λ_2 of M, and the key point response function is defined as,

$$G = \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2, \tag{10}$$

point (x, y) is consider as a key point if $G > 0$, more details can be found in [38].

The number of key points is set to 5 to 100. Then Lucas-Kanade (L-K) optical flow algorithm [39] is applied to track these key points, with forward-backward (F-B) error [39] employed to evaluate the matching accuracy of key points between two consecutive frames.

Key points with F-B error less than a preset threshold are regarded as successful tracking points. If the average spatial displacement of all successful tracking points is less than 0.5 pixel across 5 consecutive frames, we consider the background scene is static without camera jitters.

3.2.3. Change Detection Based CFs for Three-Stage Re-Detection

Based on the accurate motion estimation of background, change detection based correlation filter (CDCF) tracker is further proposed to take advantage of target's motion features. CDCF is coupled with SiamRCNN's two-stage re-detection to form a three-stage re-detection framework. The pipeline of CDCF module is shown in Figure 2. The red area on the left side indicates the Gaussian Mixture Model (GMM) construction process, which is used to describe the background and updated every frame. The blue area shows the computation process of correlation filter, and the green area depicts the combination of motion features and the correlation filter tracker to finally obtain a credible target state output.

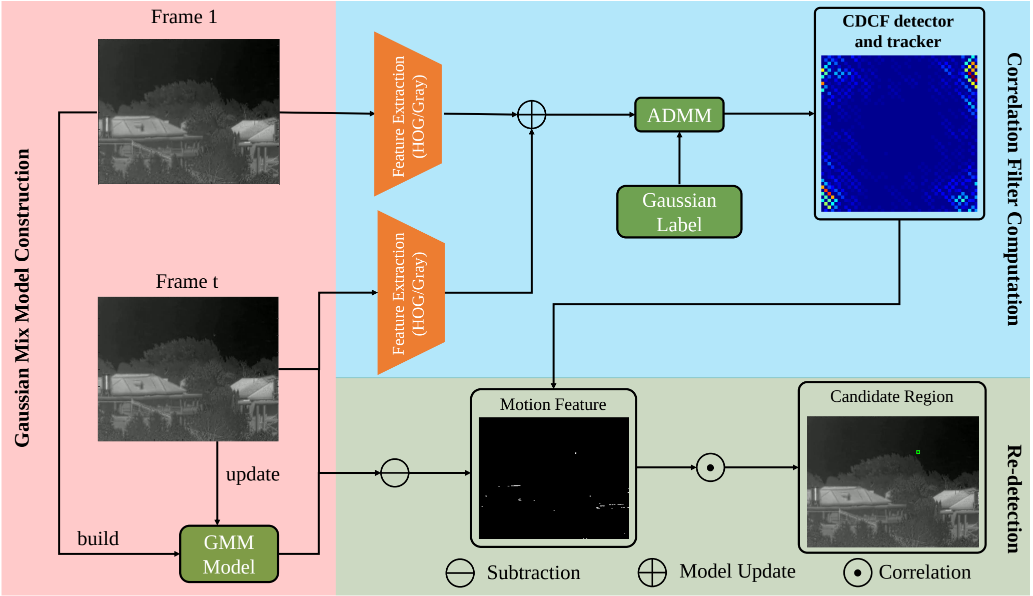
Gaussian Mix Model Construction
Frame 1
Frame t
update
build
GMM Model
Feature Extraction (HOG/Gray)
Feature Extraction (HOG/Gray)
ADMM
Gaussian Label
CDCF detector and tracker
Correlation Filter Computation
Motion Feature
Candidate Region
Re-detection
Subtraction
Model Update
Correlation

Figure 2. Pipeline of CDCF. We use the first frame of sequence to construct a Gaussian Mix Model (GMM), and update GMM model every frame. CDCF extracts hand-drafted feature to build a correlation filter, which can track target and redetect target when target lost.

Background Modeling. When the background is static, each pixel is normally distributed in the time domain, pixels within a certain threshold are judged as background and vice versa as moving targets. Based on this assumption, a Gaussian mixed model (GMM) is built to capture moving targets. Denote X_t as the intensity value of pixel (x, y) in frame t, and the GMM model is computed as,

$$P(X_t) = \sum_{i=1}^{K} \kappa_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}), \tag{11}$$

where K is the number of Gaussian components (usually ranges from 3 to 5), $\kappa_{i,t}$ is the weight of component i in frame t, $\mu_{i,t}$ and $\Sigma_{i,t}$ are the mean and variance matrix of component i, respectively. Gaussian probability density function $\eta(X_t, \mu_{i,t}, \Sigma_{i,t})$ is defined as,

$$\eta(X_t, \mu_{i,t}, \Sigma_{i,t}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t-\mu_t)^T \Sigma^{-1} (X_t-\mu_t)}. \tag{12}$$

For a pixel value, X_t, it will be checked from the existing K Gaussian components, until a match is found. The match is defined as success if the pixel value X_t is within 2.5 times the standard deviations of a component. Then, GMM model is updated as,

$$\kappa_{i,t} = (1-\alpha)\kappa_{i,t-1} + \alpha Q_{i,t}, \tag{13}$$

where α is the learning rate, $Q_{i,t}$ equals 1 when the matching is successful and 0 otherwise. Keep the parameters μ, Σ for unmatched components unchanged, and update matched component as,

$$\mu_t = (1-\rho)\mu_{t-1} + \rho X_t, \tag{14}$$

$$\Sigma_t = (1-\rho)\Sigma_{t-1} + \rho(X_t - \mu_t)^{\mathrm{T}}(X_t - \mu_t), \tag{15}$$

where ρ is learning rate:

$$\rho = \alpha\eta(X_t|\mu_k, \sigma_k). \tag{16}$$

If X_t does not match any of the K components, we classify the pixel as motion target (As we can see in the right subfigures of Figure 3, the pink region is the visualization of motion feature, represents the presence probability of the moving target). Based on the motion feature, SiamSTA can perceive tiny moving target with little semantic information in dynamic background.

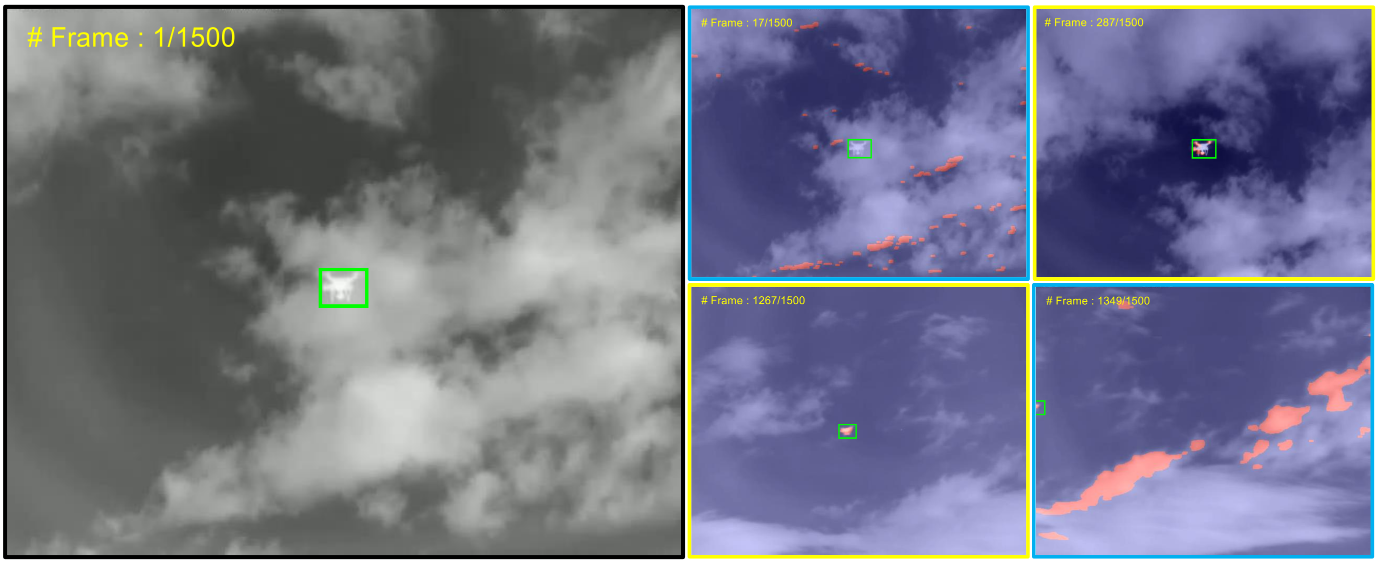

Figure 3. Motion features of CDCF. The areas with motion features are marked in pink, the rest in blue. When background is static e.g., frame 287, 1267, the motion feature is quite distinct, which is suitable for CDCF to perceive target, even the target is tiny and with little semantic information. When dynamic background occurs e.g., frame 17, 1349, the moving clouds in the scene will cause a heavy disturbance in discerning the real target.

Change Detection Correlation Filter. CDCF is trained using the initial frame and perform correlation operation to track the target in subsequent frames. Previous work [37] exploits motion feature as a re-detection method, which simply treats the motion feature as candidate region of target, and restart tracking on the new position based on the motion feature. However, motion feature is not accurate enough to describe target state. As shown in Figure 3 (frame 17), when there are moving distractors in the background and the target remains static, the motion feature will produce a strong response value in the background distractors, while the response at the target position is zero, which will easily lead to tracking drift.

To make better use of motion feature, this work proposes a novel correlation filter which is not only used as tracking, but also used as re-detection. To construct a robust correlation filter, a novelty objective function is proposed, which can perceive background distractors [40]:

$$E(f_k, \boldsymbol{H}) = \frac{1}{2}\left\|\sum_{d=1}^{D}(\boldsymbol{B}x_k^d * \boldsymbol{f}_k^d) - \boldsymbol{H}\right\|_2^2 + \frac{1}{2}\sum_{d=1}^{D}\left\|\boldsymbol{f}_k^d\right\|_2^2 + \frac{\alpha}{2}\|\boldsymbol{H} - \boldsymbol{y}\|_2^2, \tag{17}$$

where $\boldsymbol{x}_k^d, \boldsymbol{f}_k^d \in \mathbb{R}^T$ denote the d-th channel of the vectorized image and filter of frame k, respectively. D is the total channel number. $\boldsymbol{y} \in \mathbb{R}^T$ is the expected response (with Gaussian distribution). $*$ indicates convolution operator. $\boldsymbol{B} \in \mathbb{R}^{M \times N}$ is a cropping matrix to select central M elements in $\boldsymbol{x}_k^d$, N is the length of x. Usually $N >> M$, $\boldsymbol{H}$ is expected response output. α is a adaptive parameter during tracking, which can be calculated from APCE [41]:

$$\alpha = APCE = \frac{1}{WH} \cdot \frac{|F_{max} - F_{min}|^2}{\sum_{i=1}^{N}\sum_{j=1}^{H}(F_{i,j} - F_{min})^2}. \tag{18}$$

APCE is a confidence evaluation method, when the target state is quite credible, the APCE value is high, otherwise, the value will decrease. We use it as the regularization coefficient, when background is clear, APCE is very high, we make the expected output closer to the Gaussian response y, and when there are cluttered disturbances in the background, we choose to tolerate them so that the expected output has a non-zero value response at the background.

Then Alternating Direction Method of Multipliers (ADMM) algorithm is applied to minimize Equation (17) to achieve a local optimal solution. The augmented Lagrangian form of the equation can be formulated as:

$$\begin{aligned} L_k(f_k, \hat{g}_k^d, \xi^T, \hat{\boldsymbol{H}}) &= E(f_k, \hat{g}_k^d, \hat{\boldsymbol{H}}) + \hat{\xi}^T(\hat{g}_k - \sqrt{N}(\boldsymbol{F}\boldsymbol{B}^T \otimes I_D)\boldsymbol{f}_k) \\ &+ \frac{\rho}{2}\left\|\hat{g}_k - \sqrt{N}(\boldsymbol{F}\boldsymbol{B}^T \otimes I_D)\boldsymbol{f}_k\right\|_2^2, \end{aligned} \tag{19}$$

where $\hat{\xi}^T = [\hat{\xi}_1^T, \hat{\xi}_2^T, \ldots, \hat{\xi}_D^T]$ is the $1 \times DN$ Lagrangian vector in Fourier domain, ρ is a penalty factor. Then, ADMM is applied by alternately solving the following sub-problems:

$$\begin{aligned} \hat{g}_{k+1} = argmin \frac{1}{2}\left\|\sum_{d=1}^{D}(\hat{x}_k^d \odot \hat{g}_k^d) - \hat{\boldsymbol{H}}\right\|_2^2 + \hat{\xi}^T(\hat{g}_k - \sqrt{N}(\boldsymbol{F}\boldsymbol{B}^T \otimes I_D)\boldsymbol{f}_k) \\ + \frac{\rho}{2}\left\|\hat{g}_k - \sqrt{N}(\boldsymbol{F}\boldsymbol{B}^T \otimes I_D)\boldsymbol{f}_k\right\|_2^2, \end{aligned} \tag{20}$$

$$\boldsymbol{f}_{k+1} = argmin \frac{1}{2}\sum_{d=1}^{D}\left\|\boldsymbol{f}_k^d\right\|_2^2 + \hat{\xi}^T(\hat{g}_k - \sqrt{N}(\boldsymbol{F}\boldsymbol{B}^T \otimes I_D)\boldsymbol{f}_k) + \frac{\rho}{2}\left\|\hat{g}_k - \sqrt{N}(\boldsymbol{F}\boldsymbol{B}^T \otimes I_D)\boldsymbol{f}_k\right\|_2^2, \tag{21}$$

$$\hat{\boldsymbol{H}} = \underset{\hat{\boldsymbol{H}}}{argmin} \frac{1}{2}\left\|\sum_{d=1}^{D}(\hat{x}_k^d \odot \hat{g}_k^d) - \hat{\boldsymbol{H}}\right\|_2^2 + \frac{\alpha}{2}\|\hat{\boldsymbol{H}} - \hat{\boldsymbol{y}}\|_2^2, \tag{22}$$

which can be easily solved as follows:

$$\hat{g}_{k+1}(t) = \frac{1}{\rho N}\left(I - \frac{\hat{\boldsymbol{x}}(t)\hat{\boldsymbol{x}}(t)^T}{\rho N + \hat{\boldsymbol{x}}(t)\hat{\boldsymbol{x}}(t)^T}\right)(\hat{\boldsymbol{H}}(t)\hat{\boldsymbol{x}}(t) - N\xi^T + N\rho\hat{\boldsymbol{f}}_k(t)). \tag{23}$$

$$f_{k+1} = \frac{\xi^T N + \rho N \boldsymbol{g}_{k+1}}{1 + \rho N}. \tag{24}$$

$$\hat{\boldsymbol{H}} = \frac{\sum_{d=1}^{D}(\hat{g}_k^d * \hat{x}_k^d) + \alpha\hat{\boldsymbol{y}}}{1 + \alpha}. \tag{25}$$

Beside used for tracking target, the correlation filter calculated from (23) is also served as a detector in CDCF.

Then the appearance model $\hat{x}_k^{model}$ is updated as follows:

$$\hat{x}_k^{model} = (1-\eta)\hat{x}_{k-1}^{model} + \eta\hat{x}_k, \tag{26}$$

where η is the learning rate of the appearance model.

Remark 1. *To the best of our knowledge, no correlation filters have been used for re-detection. This is mainly because the correlation filter is very sensitive to the position offset, and when there is a large position deviation between the correlation filter model and the target, the correlation filter will get a very low response value, and the target will be detected only when the position deviation is very small. Therefore, how to judge the location where the target reappears is crucial for whether the correlation filter can find the target again. Thanks to the robust motion feature, CDCF will obtain a suitable target candidate region when background is static. Then, CDCF will correct the target position and scale based on correlation operation.*

3.3. Online Tracking and Updating

As aforementioned, local tracking equipped with spatial-temporal constraints helps to locate small targets with limited semantic information. Global re-detection, instead, could be more reliable when faced with challenges like occlusion and out-of-view in long-term tracking. Hence, it is crucial to learn to switch adaptively between local tracking and global re-detection to leverage their complementary strengths.

Local Tracking: Suppose $c_i = [c_{i,start}, c_{i,start+1}, \ldots, c_{i,t-1}]$ is a continuous sub-trajectory from frame $start(c_i)$, and $c_{i,t-1}$ is a trustworthy tracking result in frame $t-1$. For frame t, previous trusted predictions in $[c_1, c_2, \ldots, c_i]$ are fed into the second stage of the re-detector. For static background, only proposals with an overlap greater than 0.01 with the bounding box in $c_{i,t-1}$ are considered as target candidates. If the re-detector finds a proposal with a confidence score over 0.5, local tracking is believed to be valid, and its corresponding result $c_{i,t}$ is added to c_i. Otherwise, local tracking is paused, indicating the end of this continuous trajectory.

Global Re-detection: Starting from the failed frame, global re-detection is performed. Like SiamRCNN, SiamSTA also track potential similar-looking distractors and record their trajectories A. Then the results of CDCF is introduced to guide the global re-detection. SiamSTA compares the bounding box size S_{CD} predicted by CDCF with the target size of previous frames. When the background is static and $S_{CD} \in [S_{min}, S_{max}]$, CDCF's results are judged to be credible, initialize a new sub-trajectory c_{i+1}, and restart local tracking. Otherwise, SiamSTA treats it as a reference result to facilitate selecting the suitable proposal as output for the current frame. When the predicted bounding box of CDCF has a large overlap with a high-confidence proposal, SiamSTA considers the target has been successfully recaptured and initialize a new sub-trajectory c_{i+1}, then restart local tracking. If there is no overlap between CDCF and high-confidence proposals, SiamSTA chooses the proposal with highest score as current frame output. For dynamic background, the continuous sub-trajectory is terminated directly, and the tracker relies on global re-detection to estimate the position and size of the target.

4. Experiments

4.1. Experimental Setup

4.1.1. Evaluation Metrics

This experiment uses three widely-used metrics to evaluate, including precision plot, success rate plot and average overlap accuracy. The first metric computes the percentages of frames in which the estimated target location is within a given distance threshold to the ground-truth. The second one measures the fractions of successful frames where the Intersection over Union (IoU) between the predicted bounding box and ground-truth is greater than a certain threshold varied from 0 to 1. The last one is the evaluation metric given in the Anti-UAV benchmark [16]. It calculates the mean IoU of all videos. In this experiment, an error threshold of 20 pixels are adopted in the precision plot, and the area under the curve (AUC) of the success plot is used to evaluate tracking performance.

4.1.2. Network Parameters

Our SiamSTA is built upon SiamRCNN network, and SiamSTA also borrows its trained weights. The max corners, min distance and block size for computing the background key points are set to 500, 7 and 7 respectively. For optical flow, SiamSTA utilizes a two-level pyramid with a 15 × 15 sliding window. The F-B error threshold for selecting the correct key points is set to 1.0. If the average moving distance of these selected key points for 5 consecutive frames is less than 0.5, SiamSTA considers the background to be static. A 5 × 5 median filter is used to remove the tiny foreground noises in the change detection. The weight w_r, for the first stage of re-detection is set to 0.1, hence the weight for the second re-detection stage is 0.9. The location score weight w_l is set to 5.5. In the global detection phase, the settings are consistent with SiamRCNN. As for CDCF, the learning rate η is set as 0.02.

Remark 2. *Due to the competition restrictions of the 2nd Anti-UAV challenge, we did not perform additional training on the model when testing the Anti-UAV2020 [14] and Anti-UAV2021 [15] test-dev dataset, and directly exploit the model trained on RGB datasets given by SiamRCNN. While the Anti-UAV [16] dataset contains a training subset, SiamSTA is retrained on the train subset of the Anti-UAV dataset, with a total of 160 sequences. During training, we also apply motion blur, grayscale, gamma, flip and scale augmentations. The other parameters about training is the same as the SiamRCNN [13].*

4.1.3. Details about UAV Platform

In the evaluation datasets, the picture sizes include 640 × 512, 640 × 480 and 1280 × 720 (in pixels). The detection range (distance from UAVs) varies from 0.1 km to 2.5 km. The UAV size in the image ranges from 5 × 5 to 60 × 90 (in pixels), including a variety of UAV platform such as DJI-Phantom4 (196 × 289.5 × 289.5 mm), DJI-Mavic-Air (168 × 184 × 64 mm), DJI-Spark (143 × 143 × 55 mm), DJI-Mavic-Pro (322 × 242 × 84 mm), DJI-Inspire (438 × 451 × 301 mm) and Parrot.

Remark 3. *The evaluation videos include a variety of UAV platforms with detection distances ranging from 0.1 to 2.5 km, which fully demonstrates the generality of our approach.*

4.2. Comparing with State-of-the-Arts Trackers

Comprehensive experiments are conducted to compare our SiamSTA with some of the currently best performing deep trackers, i.e., SiamRCNN [13], SiamRPN++ [8], Globaltrack [42], PrDiMP [11], DiMP [43], ATOM [10], KYS [44], SiamRPN [27] and other recent CF trackers including AutoTrack [45], ECO [46], ARCF [47], STRCF [48], KCF [18], CSRDCF [49]. For a fair comparison, these compared algorithms are reproduced on our platform with their default parameter settings maintained. The details of different comparison results on different datasets are listed below.

Anti-UAV2020 test-dev datasets [14]: The datasets contains 100 high quality IR videos and 100 RGB videos, spanning multiple occurrences of multi-scale UAVs with complex backgrounds such as clouds, urban buildings, etc. The results of the precision plots and success plots which compare the trackers mentioned above on Anti-UAV2020 test-dev datasets are shown in Figure 4. It is obviously that the proposed SiamSTA can perform better than the other trackers. SiamSTA outperforms the previous best tracker SiamRCNN [13] by 2.32% and 0.97% in terms of precision and success, respectively.

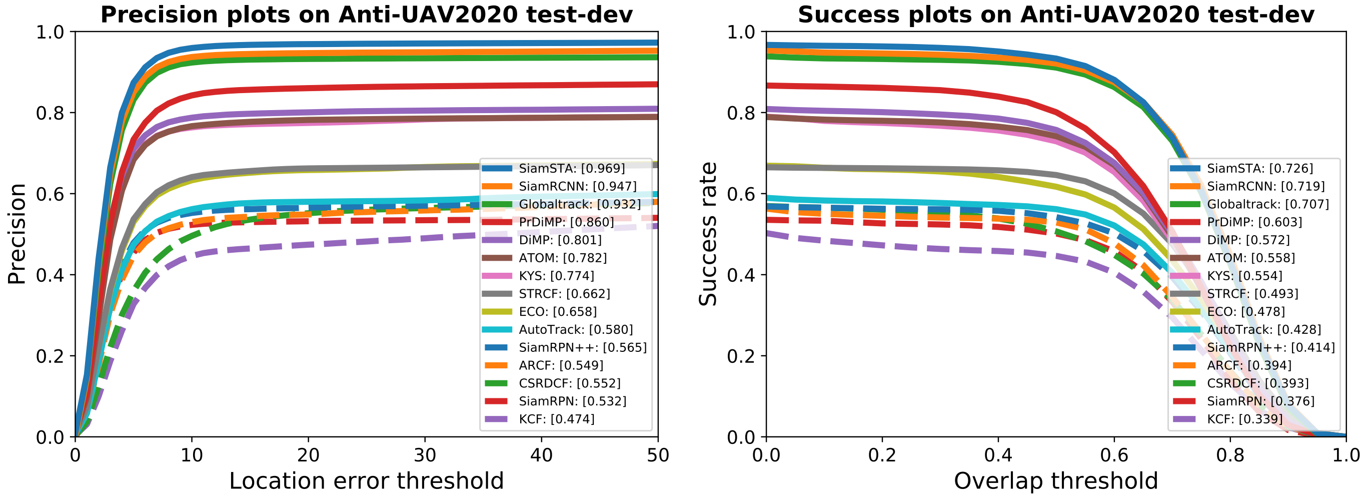

Figure 4. Precision and success plots of our SiamSTA and state-of-the-art trackers on the Anti-UAV2020 test-dev dataset. The mean precision and AUC scores are reported for each tracker. Best viewed in color and zoom.

Anti-UAV2021 test-dev datasets [15]: Based on Anti-UAV2020 test-dev, the 2021 version [15] abandons RGB videos and extends the IR data of the former to 140 videos. Furthermore, the dataset incorporates more complex scenarios such as sea, forest, mountain, and more challenging issues such as tiny objects, weak targets, which makes the tracker easily overwhelmed in the clustered backgrounds. Figure 5 reports the comparison results on Anti-UAV2021 test-dev. As one can see, SiamSTA yields the best precision score 0.888, which surpasses the second-best (SiamRCNN [13]) and third-best (GlobalTrack [42]) trackers by 6.09% and 9.09%, respectively. What is more, SiamSTA is also the best tracker in terms of success with a score of 0.655. Notably, the performance gains of our algorithm in the Anti-UAV2021 test-dev dataset are more impressive than that of Anti-UAV2020. This is mainly because the Anti-UAV2021 test-dev introduces many tiny and weak target videos, while our spatio-temporal attention and change detection are exactly designed to address such challenges, thus leading to a higher accuracy.

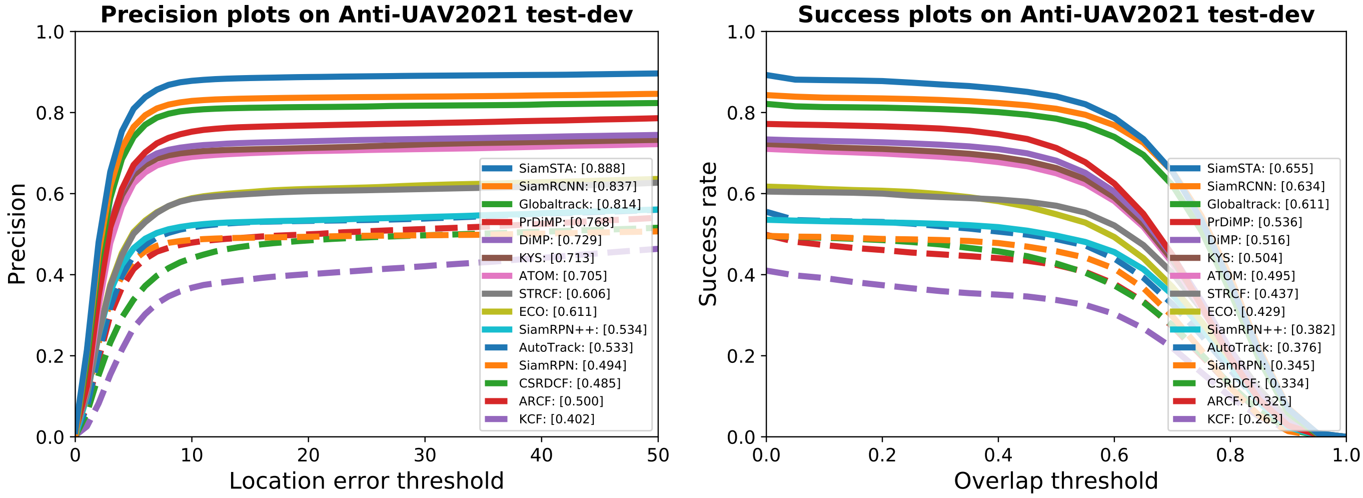

Figure 5. Precision and success plots of our SiamSTA and state-of-the-art trackers on the Anti-UAV2021 test-dev dataset. The mean precision and AUC scores are reported for each tracker. Best viewed in color and zoom.

Anti-UAV [16]: The dataset contains 318 high quality thermal infrared sequences, including train (160 sequences), test (91 sequences) and validation (67 sequences) sub-datasets. Besides the 14 trackers mentioned above, this experiment also introduce 15 state-of-the-art trackers for comparison, including, KeepTrack [50], Stark [51], HiFT [52], STMTrack [53], TransT [54], TransformerTrack [55], ROAM [56], SiamBAN [57], Siam-CAR [58], SiamFC++ [59], TADT [60], DaSiamRPN [61], BACF [21], SiamFC [6], SRDCF [20] and DSST [23]. The results are shown in Table 1, SiamSTA performs best in both test and validation sets, suppressing the previous best tracker SiamRCNN by 4.76% and 1.81%, respectively.

Table 1. The overall average accuracy (%) of state-of-the-art trackers on Anti-UAV test and validation sets. The top three results are highlighted in red, green and blue, respectively.

Method	Source	Test	Validation
DSST [23]	BMVC14	33.11	39.86
KCF [18]	T-PAMI15	33.33	38.53
SRDCF [20]	ICCV15	41.00	46.99
SiamFC [6]	ECCVW16	37.51	45.14
BACF [21]	ICCV17	40.94	45.74
ECO [46]	CVPR17	43.68	51.95
STRCF [48]	CVPR18	44.89	50.63
SiamRPN [27]	CVPR18	41.64	43.87
DaSiamRPN [61]	ECCV18	39.61	44.64
ARCF [47]	ICCV19	40.55	43.82
ATOM [10]	CVPR19	49.98	59.82
TADT [60]	CVPR19	43.52	55.20
SiamRPN++ [8]	CVPR19	42.58	45.88
DiMP50 [43]	ICCV19	49.33	62.48
PrDiMP50 [11]	CVPR20	55.70	62.61
AutoTrack [45]	CVPR20	38.70	47.49
SiamFC++ [59]	AAAI20	44.92	50.44
KYS [44]	ECCV20	46.70	60.35
GlobalTrack [42]	AAAI20	**64.31**	**73.84**
SiamCAR [58]	CVPR20	46.59	54.79
SiamBAN [57]	CVPR20	39.53	42.42
Siam R-CNN [13]	CVPR20	**65.16**	**74.76**
ROAM [56]	CVPR20	45.15	56.15
TransformerTrack [55]	CVPR21	54.75	65.21
TransT [54]	CVPR21	52.14	60.86
STMTrack [53]	CVPR21	40.86	46.41
HiFT [52]	ICCV21	37.87	47.41
Stark [51]	ICCV21	59.08	69.03
KeepTrack [50]	ICCV21	61.05	67.95
SiamSTA	Ours	**68.26**	**76.11**

4.3. Qualitative Evaluation

Figure 6 shows qualitative comparisons between SiamSTA and other state-of-the-art trackers. SiamSTA shows clear superiority over other trackers in handing challenging tracking situations, including Motion Blur, Fast Motion, Thermal Crossover, Out-of-view and Occlusion. With the help of the proposed spatio-temporal attention mechanism and change detection correlation filter, SiamSTA performs excellent against target loss and thermal crossover, which are common in TIR tracking.

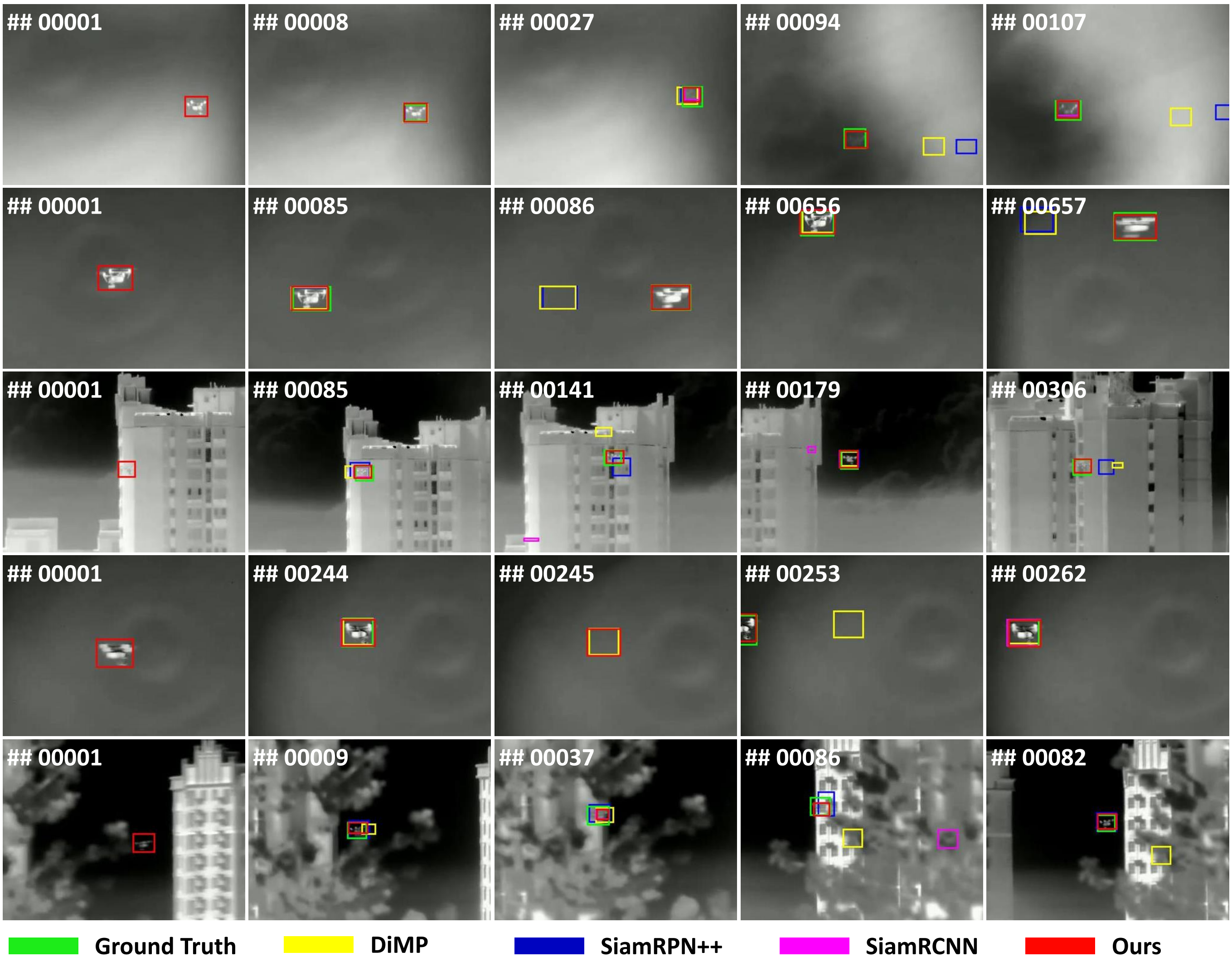

Figure 6. Qualitative comparison of SiamSTA with other state-of-the-art trackers in handling different challenging scenarios. From top to bottom are Motion Blur (Sequence 20190925_222534_1_7_1), Fast Motion (Sequence 20190925_210802_1_3_1), Thermal Crossover (Sequence 20190925_133630_1_9_1), Out-of-view (Sequence 20190925_210802_1_8_1) and Occlusion (Sequence 20190925_143900_1_5_1). The tracking video can be found in https://youtu.be/_l4hP1ZWG3w (accessed on 17 March 2022).

4.4. Attribute-Based Evaluation

For a comprehensive evaluation of SiamSTA, in addition to the 5 attributes (Thermal crossover, scale variation, out-of-view, fast motion and occlusion) defined in Anti-UAV [16], this experiment has added a new attribute, tiny scale (with a diagonal length less than 10 pixels). Figure 7 reports the comparison results (precision and success plots) of SiamSTA and other state-of-the-art trackers in different attributes. It can be seen that SiamSTA achieves the highest scores on all attributes, which fully validates the superior performance of SiamSTA in tackling the challenges of various attributes. Especially noteworthy is the fact that SiamSTA performs particularly well in Thermal Crossover and Tiny Scale attributes, surpassing the second-best tracker SiamRCNN by 7.76% and 7.45% in terms of precision, 7.24% and 9.51% in terms of success rate, respectively.

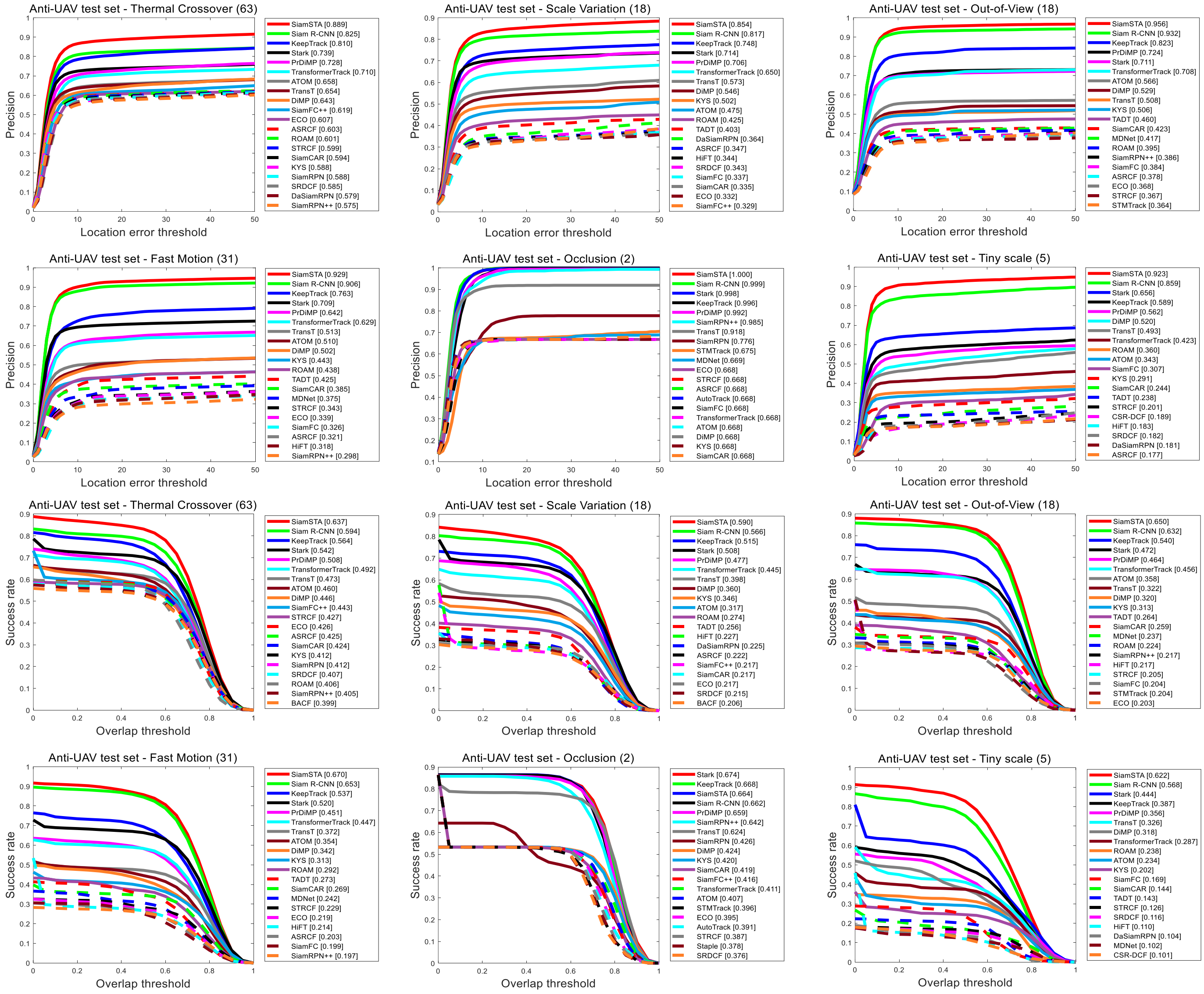

Figure 7. Precision and success plots of our SiamSTA and state-of-the-art trackers on the different attributes defined in Anti-UAV [16] test set. The mean precision and AUC scores are reported for each tracker. Best viewed in color and zoom.

4.5. Tracker Robustness Testing against Weather Challenges

Weather conditions have a critical impact on tracker performance, especially for anti-UAV tracking in real-world scenarios where complex weather conditions such as foggy days, cloudy days, and nighttime bring new challenges to tracking. For this purpose, we selected video sequences from the Anti-UAV [16] dataset under different weather conditions for visualization and analysis. As shown in the Figure 8, the left side is the scene captured by the visible camera, and the right side is the infrared imaging scene at the same moment and under the same shooting angle. It can be seen that in extreme weather conditions, the infrared imaging quality is superior to the visible imaging quality and thus allows a clearer view of the UAV target, so the infrared anti-UAV tracking is more adaptable to the environment. At the same time, SiamSTA can still stably track the target under complex weather conditions, providing a strong guarantee for anti-UAV tracking.

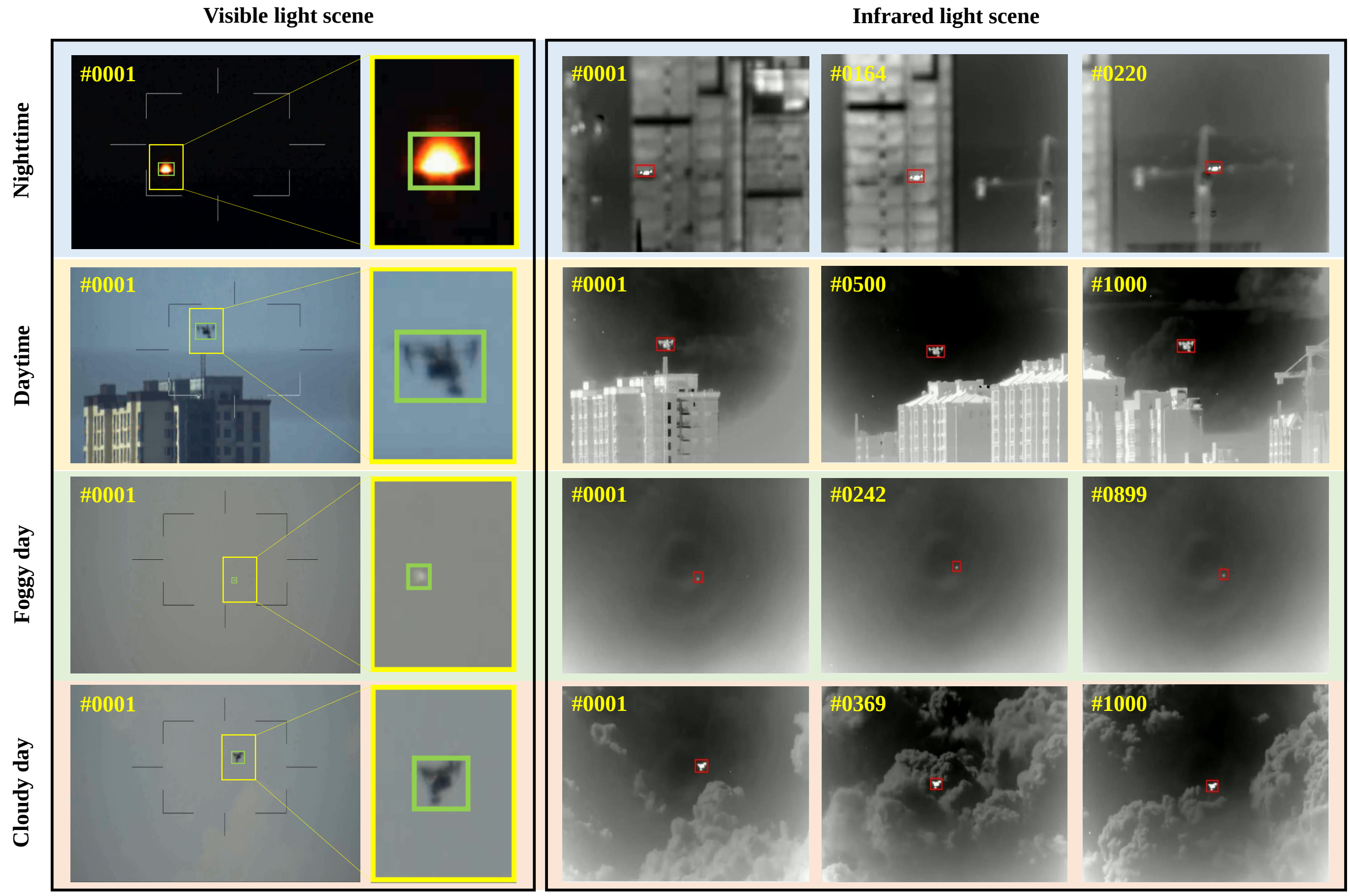

Figure 8. Visual analysis of SiamSTA for different weather challenges. From top to bottom are Sequence 20190926_195921_1_9, Sequence 20190925_101846_1_8, Sequence 20190925_133630_1_2, Sequence 20190925_131530_1_4. The red rectangle indicates the predicted results of SiamSTA.

4.6. Ablation Study

We perform an ablation study to demonstrate the impact of each component in the proposed SiamSTA method on Anti-UAV2021 test-dev [15]. Average tracking accuracy defined in Anti-UAV [16] is adpoted as the evaluation criteria. The baseline method is the original SiamRCNN [13] method.

Effects of Lost Estimation. SiamSTA treats the target state as lost when the confidence score falls below 0. As shown in Table 2, integrating lost estimation brings an improvement of 0.71% over the SiamRCNN baseline, validating that this simple operation is quite effective.

Table 2. Ablation studies on components of SiamSTA. Lost: lost estimation, STA: spatio-temporal attention, CD: change detection.

	Lost	STA	CD	Score (%)
Baseline				64.29
	✓			64.70
		✓		65.61
			✓	66.44
	✓	✓	✓	67.30
GLCF				37.01
		✓	✓	56.04

Effects of Spatio-Temporal Attention. To verify the effect of Spatio-Temporal attention, a variant is created by adding spatio-temporal attention (STA) to baseline. Result in Table 2 shows the effectiveness of Spatio-Temporal Attention (STA) that leads to 2.05% improvement in average tracking accuracy. This can be attributed to the precise switching between global re-detection and nearby tracking, which suppresses the disturbance of cluttered background and thus improves the robustness of tracking.

Effects of Change Detection. We further explore the effectiveness of Change detection (CD). Through purely adding CD to the baseline, the tracking result achieves a performance lift of 2.15% (from 64.29% to 66.44%), the best among all three components, which can be mainly credited to the precise perception ability of tiny and weak target.

To further demonstrate the universality of our approach, we incorporate CD and STA into GLCF tracker [40], which achieves a score of 56.04%, a 51.42% performance improvement over the original GLCF tracker. This indicates that motion feature used in CDCF is generic and applicable for various TIR trackers.

5. Conclusions

This paper proposes a novel algorithm called SiamSTA, which fully exploits the prior knowledge to inspire the current tracker to make optimal decisions. SiamSTA first employs a spatio-temporal attention mechanism to limit the candidate proposals focus on the validate regions and reduce the interference caused by background distractors. Then a CDCF re-detection submodule is introduced into SiamSTA to combat the challenges of target occlusion and out of view. Finally, SiamSTA achieves high-precision online tracking and high-confidence feedback updates by combining local search and global detection. Extensive experiments on three anti-UAV datasets have demonstrated the effectiveness of our SiamSTA, and we strongly believe that our work can promote the development of visual tracking in remote sensing and its application in anti-UAV missions.

Author Contributions: Conceptualization, J.C. and B.H.; methodology, J.C. and B.H.; software, J.C. and B.H.; validation, J.C., B.H. and J.L..; formal analysis, J.L. and T.X.; investigation, J.C., B.H., Y.W. and M.R.; resources, J.L. and T.X.; data curation, J.L. and T.X.; writing—original draft preparation, J.C. and B.H.; writing—review and editing, J.L., Y.W., M.R. and T.X.; visualization, J.C., B.H., Y.W. and M.R.; supervision, J.L. and T.X.; project administration, J.L. and T.X.; funding acquisition, J.L. and T.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Key Laboratory Foundation under Grant TCGZ2020C004 and Grant 202020429036.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this paper are publicly available and can be accessed at https://anti-uav.github.io/dataset/ (accessed on 20 January 2022) for Anti-UAV2021 test dev, https://github.com/ucas-vg/Anti-UAV (accessed on 20 January 2022) for Anti-UAV.

Acknowledgments: This work was supported by the Key Laboratory Foundation under Grant TCGZ2020C004 and Grant 202020429036.

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

References

1. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In Proceedings of the Computer Vision—ECCV 2016—14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
2. Fu, C.; Lin, F.; Li, Y.; Chen, G. Correlation filter-based visual tracking for uav with online multi-feature learning. *Remote Sens.* **2019**, *11*, 549. [CrossRef]
3. Xue, X.; Li, Y.; Dong, H.; Shen, Q. Robust correlation tracking for UAV videos via feature fusion and saliency proposals. *Remote Sens.* **2018**, *10*, 1644. [CrossRef]
4. Huang, B.; Xu, T.; Jiang, S.; Chen, Y.; Bai, Y. Robust visual tracking via constrained multi-kernel correlation filters. *IEEE Trans. Multimed.* **2020**, *22*, 2820–2832. [CrossRef]

5. Cliff, O.M.; Saunders, D.L.; Fitch, R. Robotic ecology: Tracking small dynamic animals with an autonomous aerial vehicle. *Sci. Robot.* **2018**, *3*, eaat8409. [CrossRef] [PubMed]
6. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016.
7. 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 Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019.
8. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019.
9. Guo, D.; Shao, Y.; Cui, Y.; Wang, Z.; Zhang, L.; Shen, C. Graph Attention Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021.
10. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ATOM: Accurate Tracking by Overlap Maximization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019.
11. Danelljan, M.; Gool, L.V.; Timofte, R. Probabilistic Regression for Visual Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
12. Huang, B.; Xu, T.; Shen, Z.; Jiang, S.; Zhao, B.; Bian, Z. SiamATL: Online Update of Siamese Tracking Network via Attentional Transfer Learning. *IEEE Trans. Cybern.* **2021**, 1–14. [CrossRef] [PubMed]
13. Voigtlaender, P.; Luiten, J.; Torr, P.H.; Leibe, B. Siam R-CNN: Visual Tracking by Re-Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
14. Anti-UAV Challenge Dataset. Available online: https://anti-uav.github.io/ (accessed on 1 May 2021).
15. Zhao, J.; Wang, G.; Li, J.; Jin, L.; Fan, N.; Wang, M.; Wang, X.; Yong, T.; Deng, Y.; Guo, Y.; et al. The 2nd Anti-UAV Workshop & Challenge: Methods and Results. *arXiv* **2021**, arXiv:2108.09909.
16. Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Guo, G.; Zhao, J.; Han, Z. Anti-UAV: A Large Multi-Modal Benchmark for UAV Tracking. *arXiv* **2021**, arXiv:2101.08466.
17. Bolme, D.; Beveridge, J.; Draper, B.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010.
18. 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]
19. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J.P. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In Proceedings of the Computer Vision—ECCV 2012—12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012.
20. Danelljan, M.; Häger, G.; Khan, F.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015.
21. Galoogahi, H.K.; Fagg, A.; Lucey, S. Learning Background-Aware Correlation Filters for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017.
22. Li, Y.; Zhu, J. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In Proceedings of the Computer Vision—ECCV 2014 Workshops, Zurich, Switzerland, 6–7 and 12 September 2014.
23. Danelljan, M.; Häger, G.; Khan, F.; Felsberg, M. Accurate scale estimation for robust visual tracking. In Proceedings of the British Machine Vision Conference, BMVC 2014, Nottingham, UK, 1–5 September 2014.
24. Li, F.; Yao, Y.; Li, P.; Zhang, D.; Zuo, W.; Yang, M.H. Integrating Boundary and Center Correlation Filters for Visual Tracking with Aspect Ratio Variation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, 22–29 October 2017.
25. Danelljan, M.; Khan, F.; Felsberg, M.; van de Weijer, J. Adaptive Color Attributes for Real-Time Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014.
26. Ma, C.; Huang, J.B.; Yang, X.; Yang, M.H. Hierarchical Convolutional Features for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015.
27. 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, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018.
28. Wan, M.; Gu, G.; Qian, W.; Ren, K.; Chen, Q.; Zhang, H.; Maldague, X. Total variation regularization term-based low-rank and sparse matrix representation model for infrared moving target tracking. *Remote Sens.* **2018**, *10*, 510. [CrossRef]
29. Zingoni, A.; Diani, M.; Corsini, G. A flexible algorithm for detecting challenging moving objects in real-time within IR video sequences. *Remote Sens.* **2017**, *9*, 1128. [CrossRef]
30. Wan, M.; Gu, G.; Qian, W.; Ren, K.; Chen, Q.; Maldague, X. Infrared image enhancement using adaptive histogram partition and brightness correction. *Remote Sens.* **2018**, *10*, 682. [CrossRef]
31. Zhang, L.; Gonzalez-Garcia, A.; Van De Weijer, J.; Danelljan, M.; Khan, F.S. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Trans. Image Process.* **2018**, *28*, 1837–1850. [CrossRef] [PubMed]

32. Felsberg, M.; Berg, A.; Hager, G.; Ahlberg, J.; Kristan, M.; Matas, J.; Leonardis, A.; Cehovin, L.; Fernandez, G.; Vojir, T. The Thermal Infrared Visual Object Tracking VOT-TIR2015 Challenge Results. In Proceedings of the IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, 7–13 December 2015.
33. Cao, Y.; Wang, G.; Yan, D.; Zhao, Z. Two algorithms for the detection and tracking of moving vehicle targets in aerial infrared image sequences. *Remote Sens.* **2016**, *8*, 28. [CrossRef]
34. Yu, X.; Yu, Q. Online structural learning with dense samples and a weighting kernel. *Pattern Recognit. Lett.* **2017**, *105*, 59–66. [CrossRef]
35. Li, M.; Peng, L.; Yingpin, C.; Huang, S.; Qin, F.; Peng, Z. Mask Sparse Representation Based on Semantic Features for Thermal Infrared Target Tracking. *Remote Sens.* **2019**, *11*, 1967. [CrossRef]
36. Wu, S.; Zhang, K.; Li, S.; Yan, J. Learning to Track Aircraft in Infrared Imagery. *Remote Sens.* **2020**, *12*, 3995. [CrossRef]
37. Huang, B.; Chen, J.; Xu, T.; Wang, Y.; Jiang, S.; Wang, Y.; Wang, L.; Li, J. SiamSTA: Spatio-Temporal Attention based Siamese Tracker for Tracking UAVs. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, 11–17 October 2021.
38. Shi, J.; Tomasi, G. Good features to track. In Proceedings of the Conference on Computer Vision and Pattern Recognition, CVPR 1994, Seattle, WA, USA, 21–23 June 1994.
39. Kalal, Z.; Mikolajczyk, K.; Matas, J. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1409–1422. [CrossRef] [PubMed]
40. Chen, J.; Xu, T.; Li, J.; Wang, L.; Wang, Y.; Li, X. Adaptive Gaussian-Like Response Correlation Filter for UAV Tracking. In Proceedings of the Image and Graphics—11th International Conference, ICIG 2021, Haikou, China, 6–8 August 2021.
41. Wang, M.; Liu, Y.; Huang, Z. Large Margin Object Tracking with Circulant Feature Maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.
42. Huang, L.; Zhao, X.; Huang, K. GlobalTrack: A Simple and Strong Baseline for Long-Term Tracking. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, 7–12 February 2020.
43. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning Discriminative Model Prediction for Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019.
44. Bhat, G.; Danelljan, M.; Van Gool, L.; Timofte, R. Know Your Surroundings: Exploiting Scene Information for Object Tracking. In Proceedings of the Computer Vision—ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020.
45. Li, Y.; Fu, C.; Ding, F.; Huang, Z.; Lu, G. AutoTrack: Towards High-Performance Visual Tracking for UAV with Automatic Spatio-Temporal Regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
46. Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.
47. Huang, Z.; Fu, C.; Li, Y.; Lin, F.; Lu, P. Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019.
48. 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, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018.
49. Lukežič, A.; Vojíř, T.; Čehovin Zajc, L.; Matas, J.; Kristan, M. Discriminative Correlation Filter with Channel and Spatial Reliability. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.
50. Mayer, C.; Danelljan, M.; Paudel, D.P.; Van Gool, L. Learning Target Candidate Association to Keep Track of What Not to Track. *arXiv* **2021**, arXiv:2103.16556.
51. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning spatio-temporal transformer for visual tracking. *arXiv* **2021**, arXiv:2103.17154.
52. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Li, Y. HiFT: Hierarchical Feature Transformer for Aerial Tracking. *arXiv* **2021**, arXiv:2108.00202.
53. Fu, Z.; Liu, Q.; Fu, Z.; Wang, Y. STMTrack: Template-Free Visual Tracking With Space-Time Memory Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021.
54. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021.
55. Wang, N.; Zhou, W.; Wang, J.; Li, H. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021.
56. Yang, T.; Xu, P.; Hu, R.; Chai, H.; Chan, A.B. ROAM: Recurrently optimizing tracking model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
57. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese box adaptive network for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
58. Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; Chen, S. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.

59. Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; Yu, G. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, 7–12 February 2020.
60. Li, X.; Ma, C.; Wu, B.; He, Z.; Yang, M.H. Target-Aware Deep Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019.
61. Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-Aware Siamese Networks for Visual Object Tracking. In Proceedings of the Computer Vision—ECCV 201—15th European Conference, Munich, Germany, 8–14 September 2018.