SiamSTA: Spatio-Temporal Attention based Siamese Tracker for Tracking UAVs

Bo Huang[1,2,†] Junjie Chen[1,†] Tingfa Xu[1,2,*] Ying Wang[1]
Shenwang Jiang[1] Yuncheng Wang[1] Lei Wang[1] Jianan Li[1,*]
[1]Beijing Institute of Technology (BIT), Beijing, China
[2]Beijing Institute of Technology Chongqing Innovation Center (BITCQIC), Chongqing, China

Abstract

With the growing threat of unmanned aerial vehicle (UAV) intrusion, anti-UAV techniques are becoming increasingly demanding. Object tracking, especially in thermal infrared (TIR) videos, though provides a promising solution, struggles with challenges like small scale and fast movement that commonly occur in anti-UAV scenarios. To mitigate this, we propose a simple yet effective spatio-temporal attention based Siamese network, dubbed SiamSTA, to track UAV robustly by performing reliable local tracking and wide-range re-detection alternatively. Concretely, tracking is carried out by posing spatial and temporal constraints on generating candidate proposals within local neighborhoods, hence eliminating background distractors to better perceive small targets. Complementarily, in case of target lost from local regions due to fast movement, a three-stage re-detection mechanism is introduced to re-detect targets from a global view by exploiting valuable motion cues through a correlation filter based on change detection. Finally, a state-aware switching policy is adopted to adaptively integrate local tracking and global re-detection and take their complementary strengths for robust tracking. Extensive experiments on the 1st and 2nd anti-UAV datasets well demonstrate the superiority of SiamSTA over other competing counterparts. Notably, SiamSTA is the foundation of the 1st-place winning entry in the 2nd Anti-UAV Challenge.

1. Introduction

Recently, unmanned aerial vehicles (UAVs) technique have received increasing attention for their wide range of applications, such as aerial photography [26], video surveillance [10, 11], and biological monitoring [4]. On the contrary, the potential abuses of this technique could lead to significant negative impacts on social 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.

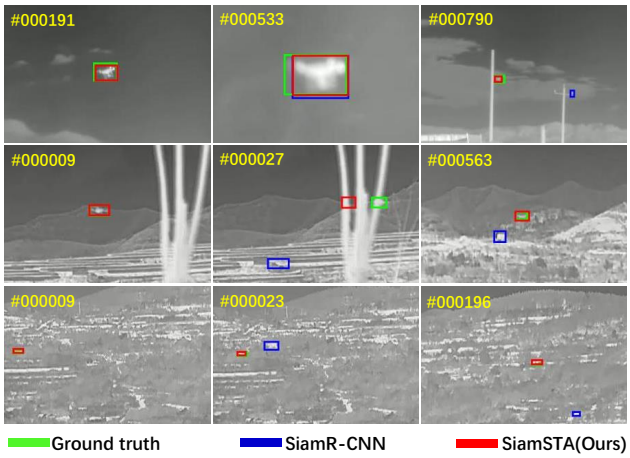

Figure 1. Qualitative comparisons of SiamSTA with SiamR-CNN baseline on three challenging sequences, with large scale variation, sever occlusion and dynamic background, respectively. Conventional SiamR-CNN tends to drift from the target in the presence of background distractors. In contrast, our SiamSTA is not prone to drift, thanks to the novel spatio-temporal attention and change detection mechanism, and thus demonstrates strong robustness in various challenging tracking scenarios.

With similar objective, visual object tracking, especially in thermal infrared (TIR) mode, servers as a fundamental step for anti-UAV task, immensely paving the way for subsequent research. As is obvious, 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 natural scenes, further introduces significant challenges such as small object, fast motion, and terminal background, etc. How to tackle

†Both authors contributed equally.
*Tingfa Xu (ciom_xtf1@bit.edu.cn) and Jianan Li (lijianan@bit.edu.cn) are the corresponding authors.

these problems remain challenging and ill-solved.

Siamese-based algorithms play a dominant role in the field of visual object tracking. SiamFC [1] 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 aspect of adding auxiliary branch [30, 12], digging deeper feature [18], improving embedding strategy [9], augmenting online mechanism [5, 7, 13], etc. While considering our tracking object as UAVs, which contain a wide range of fast motion and out-of-view situations, we argue that, unlike the improvements mentioned above, the global re-detection mechanism and trajectory modeling could play critical roles in accurate tracking.

SiamR-CNN [29] introduce global re-detection mechanism into Siamese networks and achieve outstanding tracking performance, which is thus used as our baseline algorithm. But on the other hand, as the target themselves severely lack semantic features, continuous detection mechanism could be more likely to induce tracking drift, especially when the UAV targets are drowned in terminal 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, yet two strategies that are needed in different situations. For this reason, we elaborately design a framework adaptively switching these two strategies, achieving robust tracking through performing reliable designed local tracking and wide-range re-detection alternatively.

In this paper, we propose a simple yet effective spatio-temporal attention-based Siamese network, SiamSTA, to track UAVs in thermal infrared (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 as input a search image and selects target candidates from redundant 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, the global re-detection mechanism re-detects the target by three stages: i) provide re-detections of the first-frame template, ii) implement re-detections of high-confidence predictions from previous frames, and iii) adopt 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.

We test the proposed SiamSTA on the challenging UAV infrared tracking datasets [16]. Extensive experiments show that our method is superior in handling the key challenges faced with anti-UAV tracking, including but not limited to small scale and fast movement, compared to other competing counterparts. 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:

- 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.
- 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 search for fast-moving targets in a wide range by exploiting valuable motion cues.
- Our method establishes state-of-the-art performance in the 2nd anti-UAV Challenge.

2. Related Works

In this section, we first review the development of Siamese tracking networks and then introduce some representative tracking algorithms in thermal infrared videos.

2.1. 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. [1] 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. [19] considerably enhance the tracking performance of SiamFC by introducing a region proposal network (RPN), which allows to estimate the target location, size, and ratio with the enumeration of multiple anchors. However, these trackers implement nearby search, which is difficult to recapture target after it lost. Recently, Voigtlaender et al. [29] unleash the full power of global searching by

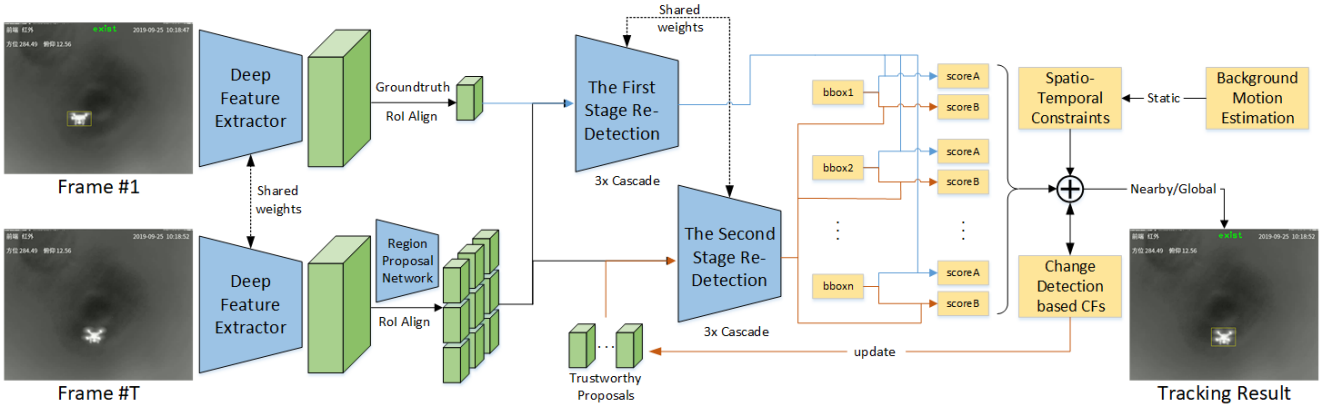

Figure 2. Overall architecture of SiamSTA. It consists of a Siamese backbone that extracts deep features from the template and the search image, followed by a three-stage re-detector that first re-detects the first-frame template, then re-locates historical predictions from previous frames, and finally fixes potential tracking failures using a change detection based CF. The symbol $\bigoplus$ 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.

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. [29] 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.

2.2. TIR Tracking

Recently, more attentions 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 image, how to extract effective features is crucial in distinguishing the target from background. [24, 8] compute motion features by thresholding the absolute difference between the current and the previous frame in pixel-wise as an extra feature channel. [32] 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 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 [31, 21, 22] start to design networks specialized for TIR images. Different to previous works, we first extract motion feature using Gaussian mixed model, then we extract the HOG feature of candidate region to train a CF tracker to assist our SiamSTA, thus better perceiving small targets in TIR images.

3. Method

In this section, we first briefly review SiamR-CNN, and then elaborate on the design of our proposed SiamR-CNN that 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 SiamR-CNN

SiamR-CNN [29] is a two-stage Siamese tracking algorithm with elaborate re-detection mechanism. Its network architecture is sequentially composed of three modules: 1) A backbone feature extraction module that contains a template branch to extract ground truth feature in target region and a test branch to prepare possible RPN proposals in search region; 2) A re-detection head module that 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 object of interest and potential similar-looking distractors based on spatio-temporal cues. In the vital third module, SiamR-CNN preserves plenty of discontinuous trajectories for making the most comprehensive decisions. Suppose one tracking trajectory consists of N non-overlapping sub-trajectories, $A = (a_1, a_2, ..., a_N)$, each sub-trajectory $a_i, \forall i \in \{1, 2, ..., N-1\}$ 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 calculated 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,

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

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

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 Eq. 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}.

SiamR-CNN backs up a lot of trajectories to ensure the success rate of re-detection. But on the other hand, tracking performance could be directly degraded right by the sophisticated search mechanism, due to the severely lack of semantic target features and complex terminal background. To address such issue, finer exploitation of spatio-temporal prior knowledge is a feasible solution.

3.2. SiamSTA Framework

Inspired by SiamR-CNN, we build our SiamSTA 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 2. To deal with background distractors, several practical guidelines using spatio-temporal attention are introduced to regulate candidate proposals. We further incorporate a collaborative strategy that combines local search and global detection to facilitate online tracking.

3.2.1 Spatio-Temporal Constraints

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, we introduce a novel spatio-temporal constraint. From the spatio perspective, considering the fact that dramatic location change of the target is less likely to appear in two adjacent frames captured by a long-range static camera, we argue that reliable tracking results can be obtained by searching the target within local neighborhoods rather than by detecting it globally. From the temporal perspective, we introduce 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, we record 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. We 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. We then update the stored target state:

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

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

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

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

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, we establish a global motion estimation model to reveal dynamic change of background scenes. To be specific, we extract the ShiTomasi [28] key points from background regions and track these points to estimate the motion of background scenes. Let $I(x, y)$ denotes the

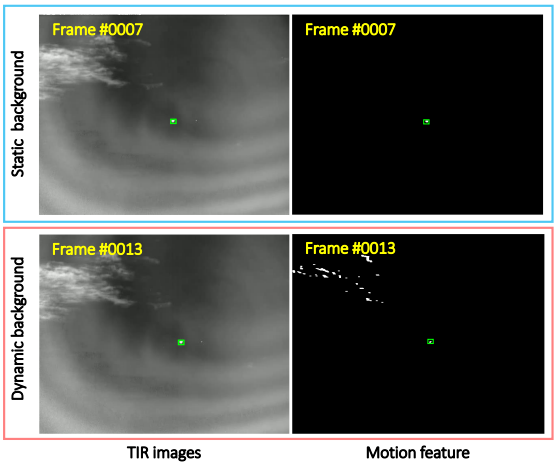

Figure 3. Motion features of CDCF. When background is static at frame #0007, the motion feature is quite distinct, which is suitable for CDCF to perceive target, even the target is tiny and with little semantic information. When dynamic background occurs at frame #0013, the moving clouds in the scene will cause a heavy disturbance in discerning the real target.

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,

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

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

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

where M is a 2×2 matrix:

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

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,

$$G = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2. \quad (10)$$

Point (x, y) is consider as a key point if $G > 0$, more details can be found in [28].

We control the number of key points in the range of 5 to 100. Then Lucas-Kanade (L-K) optical flow algorithm [17] is applied to track these key points, with forward-backward (F-B) error [17] 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

Based on the accurate motion estimation of background, we further develop a change detection based correlation filter (CDCF) tracker to take advantage of target's motion features. 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, we build a Gaussian mixed model (GMM) 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,

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

where K is the number of Gaussian components, $\kappa_{i,t}$ 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,

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

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, we update GMM model as,

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

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

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

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

where ρ is learning rate:

$$\rho = \alpha \eta(X_t | \mu_k, \sigma_k). \quad (16)$$

If X_t does not match any of the K components, we classify the pixel as motion target. Finally, we obtain accurate foreground candidate regions, as shown in Figure 3. However, candidate regions in adjacent frames are not continuous, and thus not suitable for single target tracking. So, we

introduce correlation filter (CF) to assist GMM model. We train a CF tracker using the initial frame and perform correlation operation to track the target in subsequent frames. When the background is static, we use motion feature to locate target, otherwise, we use the CF tracker to search for the target in local regions where the target last appears.

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, learn to adaptively switch between local track and global re-detection to their complementary strength is critical.

Suppose $c_i = [c_{i,start}, c_{i,start+1}, ..., c_{i,t-1}]$ 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 $[c_1, c_2, ..., c_i]$ 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.

Starting from the failed frame, global re-detection is performed. Like SiamR-CNN, we also track potential similar-looking distractors and record their trajectories A. Then we introduce the results of CDCF to guide the global re-detection. We compare the bounding box size S_{CD} predicted by CDCF with the target size of previous frames. When the background is static and $S_{CD} \in [S_{min}, S_{max}]$, we judge CDCF's results to be credible, initialize a new sub-trajectory c_{i+1}, and restart local tracking. Otherwise, we treat 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, we consider 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, we choose 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

Datasets. We use the 1st and 2nd Anti-UAV test-dev datasets to evaluate the proposed approach. The former 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 latter abandons RGB videos and extends the IR data of the former to 140 videos. Furthermore, the 2nd test-dev 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.

Evaluation metrics. We use 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.

Network parameters. Our SiamSTA is built upon SiamR-CNN network, and we also borrow 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, we utilize 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, we consider 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 SiamR-CNN. As for CDCF, we use ARCF [15] as CF tracker.

4.2. Quantitative Evaluation

We compare our SiamSTA with some of the currently best performing deep trackers, i.e. SiamRCNN [29], SiamRPN++ [18], Globaltrack [14], PrDiMP [7], DiMP [2], ATOM [5], KYS [3], SiamRPN [19] and other recent CF trackers including AutoTrack [23], ECO [6], ARCF [15], STRCF [20], KCF [27], CSRDCF [25]. For a fair comparison, these compared algorithms are reproduced on our platform with their default parameter settings maintained.

The results of the precision plots and success plots which compare the trackers mentioned above on 1st and 2nd Anti-UAV test-dev datasets are shown in Figure 4. It is obviously that the proposed SiamSTA can perform better than the other trackers. Specifically, on the precision plot of

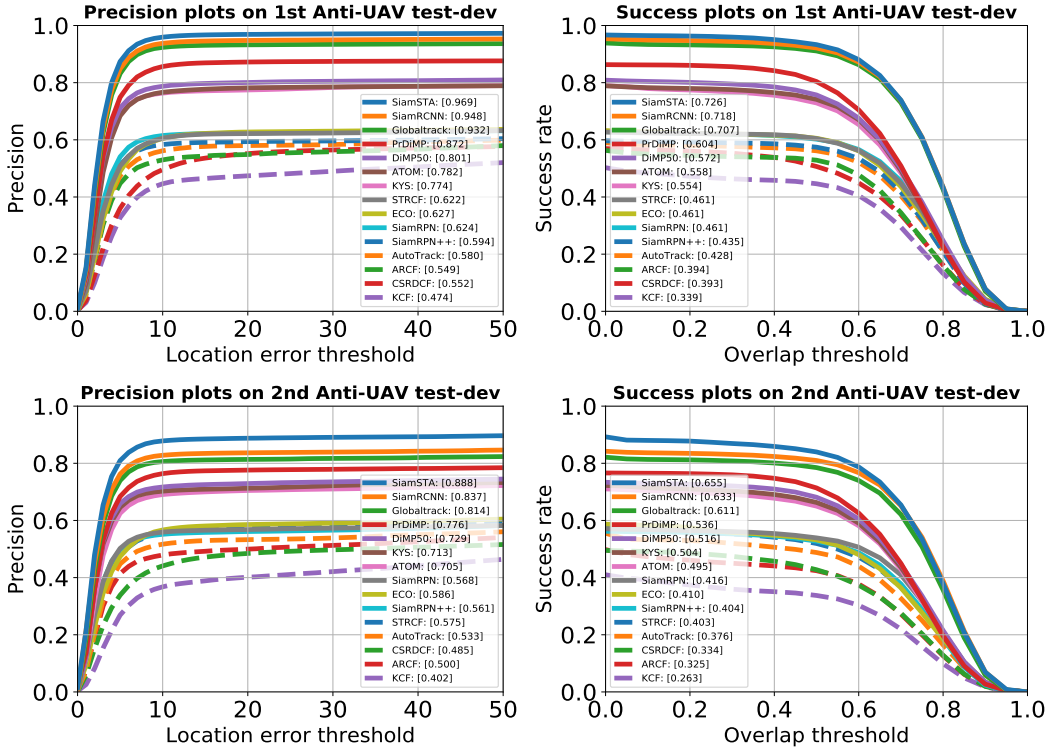

Figure 4. Precision and success plots of our SiamSTA and state-of-the-art trackers on the 1st and 2nd Anti-UAV test-dev datasets. We report the mean precision and AUC scores for each tracker. Best viewed in color.

Table 1. Comparison of average tracking accuracy (%) of our SiamSTA and other state-of-the-art trackers. The top three results are highlighted in red, green and blue, respectively.

	SiamRPN++	SiamRPN	ATOM	KYS	DiMP-50	PrDiMP	Globaltrack	SiamR-CNN	SiamSTA (Ours)
1st test-dev	44.25	46.82	56.72	56.23	58.14	62.45	72.04	72.95	74.46
2nd test-dev	41.02	42.23	50.32	51.22	52.41	55.52	62.15	64.29	67.30

the 2nd test-dev, our approach provides a mean precision score of 88.8%, outperforming the second best performing tracker, SiamR-CNN, 83.7%, by more than 5%, which is a significant improvement. Notably, the performance gains of our algorithm in the latest dataset are more impressive. This is mainly because the 2nd test-dev introduce many small 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.

Table 1 shows the overall performance of top 9 trackers in terms of the average overlap accuracy metric. Our tracker wins the first place by scoring 74.46% on the 1st dataset and 67.30% on the 2nd dataset. Moreover, our method exhibits a considerable progress over the original SiamR-CNN, especially in the 2nd test-dev, our SiamSTA obtains a gain of 4.68% in score. Moreover, our method shows a significant improvement over the original SiamR-CNN, especially in the 2nd test-dev, our SiamSTA obtained a score of 4.68%. This once again demonstrates that our spatio-temporal attention is quite effective for capturing small targets.

4.3. Qualitative Evaluation

Figure 5 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 scale variation, out-of-view, occlusion, tiny target and clustered background. Thanks to the proposed spatio-temporal attention mechanism, our SiamSTA can well track tiny targets with complex background (i.e. forest) and recapture targets quickly upon lost (i.e. out-of-view and full occlusion). In addition, SiamSTA can obtain favorable motion features based on change detection to track tiny targets confidently.

4.4. Ablation Study

We perform an ablation study to demonstrate the impact of each component in the proposed SiamSTA method on 2nd Anti-UAV test-dev. We adopt average tracking accuracy defined in Anti-UAV as the evaluation criteria. The baseline method is the original SiamR-CNN method.

Effects of Lost Estimation. We regard the target state as lost when the confidence score falls below 0. As shown

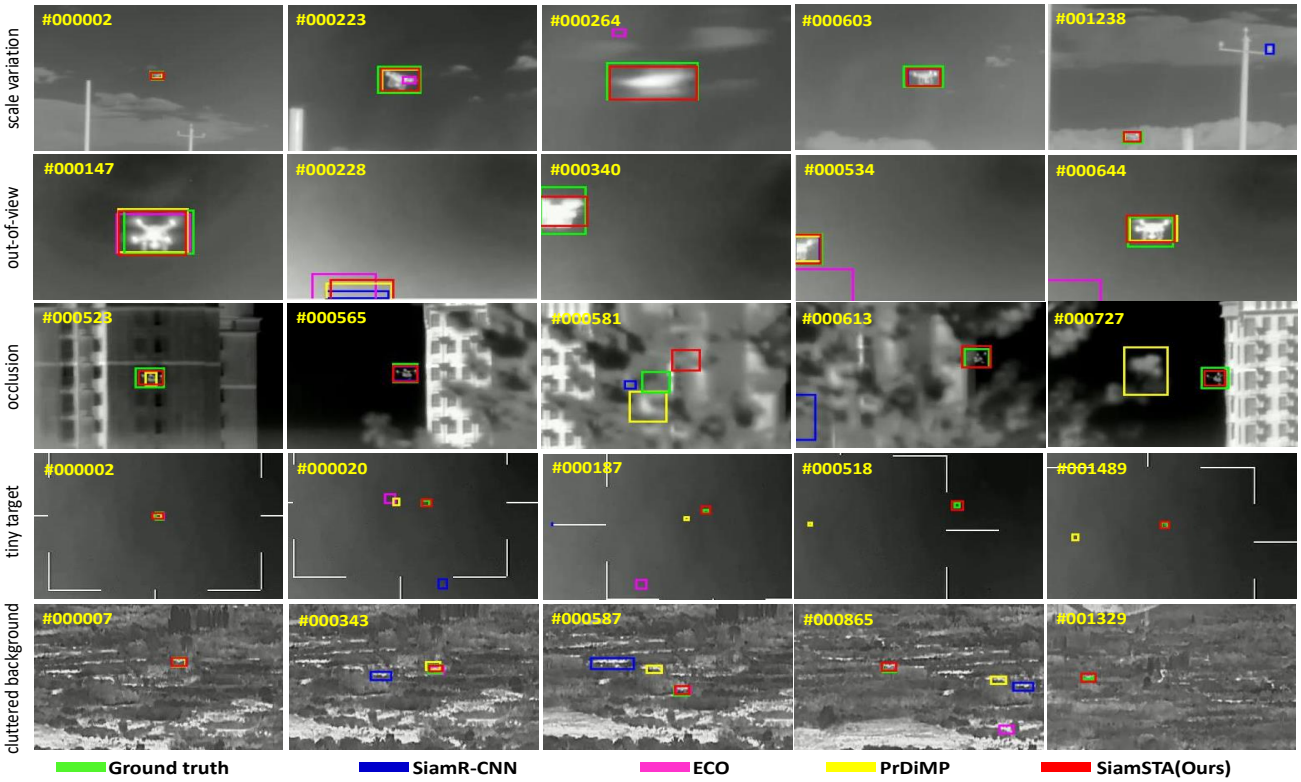

Figure 5. Qualitative comparison of SiamSTA with other state-of-the-art trackers in handling different challenging scenarios.

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**
ARCF		✓	✓	56.49

in Table 2, integrating lost estimation brings (0.71)% point improvement over the SiamR-CNN baseline, validating this simple yet effective operation.

Effects of Spatio-Temporal Attention. We create a variant which adds spatio-temporal attention (STA) to baseline. Result in Table 2 shows the effectiveness of Spatio-Temporal Attention (STA) that leads to 2.05% improvement on average tracking accuracy. This can be attributed to the precise switch between global re-detection and nearby tracking, which suppresses the disturbance of cluttered background and thus improves the tracking robust.

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 CF tracker [15], which achieves a score of 56.49%, outperforming most deep trackers listed in Table 1. This indicates that motion feature used in CDCF is generic and applicable for various TIR trackers.

5. Conclusion

In this paper, we propose a novel algorithm called SiamSTA, which fully exploits the prior knowledge to inspire the current tracker to make optimal decisions. We first employ a spatio-temporal attention mechanism to limit the candidate proposals focus on the validate regions and reduce the interference caused by background distractors. Then we introduce a CDCF re-detection submodule into SiamSTA to combat the challenges of target occlusion and out of view. Finally, we achieve high-precision online tracking and high-confidence feedback updates by combining local search and global detection. Extensive experiments on 1st & 2nd Anti-UAV have demonstrated the effectiveness of our SiamSTA, and in future work we will delve into improving the performance of SiamSTA by training on the IR data of the Anti-UAV benchmarks.

Acknowledgements. This work was supported by the Major Science Instrument Program of the National Natural Science Foundation of China under Grant 61527802, and the Key Laboratory Foundation under Grant TCGZ2020C004.

References

[1] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip H S Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016. 2

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 6

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *ECCV*, 2020. 6

[4] Oliver M Cliff, Debra L Saunders, and Robert Fitch. Robotic ecology: Tracking small dynamic animals with an autonomous aerial vehicle. *Science Robotics*, 3(23), 2018. 1

[5] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, 2019. 2, 6

[6] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, 2017. 6

[7] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020. 2, 6

[8] M. Felsberg, A. Berg, G. Hager, J. Ahlberg, M. Kristan, J. Matas, A. Leonardis, L. Cehovin, G. Fernandez, and T. Vojir. The thermal infrared visual object tracking vot-tir2015 challenge results. In *ICCV Workshops*, 2015. 3

[9] Dongyan Guo, Yanyan Shao, Ying Cui, Zhenhua Wang, Liyan Zhang, and Chunhua Shen. Graph attention tracking. In *CVPR*, 2021. 2

[10] Bo Huang, Tingfa Xu, Shenwang Jiang, Yiwen Chen, and Yu Bai. Robust visual tracking via constrained multi-kernel correlation filters. *IEEE Transactions on Multimedia*, 22(11):2820–2832, 2020. 1

[11] Bo Huang, Tingfa Xu, Jianan Li, Ziyi Shen, and Yiwen Chen. Transfer learning-based discriminative correlation filter for visual tracking. *Pattern Recognition*, 100:107157, 2020. 1

[12] Bo Huang, Tingfa Xu, Bo Liu, and Bo Yuan. Context constraint and pattern memory for long-term correlation tracking. *Neurocomputing*, 377:1–15, 2020. 2

[13] Bo Huang, Tingfa Xu, Ziyi Shen, Shenwang Jiang, Bingqing Zhao, and Ziyang Bian. Siamatl: Online update of siamese tracking network via attentional transfer learning. *IEEE Transactions on Cybernetics*, PP:1–14, 01 2021. 2

[14] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *AAAI*, 2020. 6

[15] Ziyuan Huang, C. Fu, Y. Li, Fuling Lin, and Peng Lu. Learning aberrance repressed correlation filters for real-time uav tracking. In *ICCV*, 2019. 6, 8

[16] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Guodong Guo, Jian Zhao, and Zhenjun Han. Anti-uav: A large multi-modal benchmark for uav tracking. *arXiv preprint arXiv:2101.08466*, 2021. 2, 6

[17] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2011. 5

[18] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 2, 6

[19] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 2, 6

[20] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *CVPR*, 2018. 6

[21] Meihui Li, Lingbing Peng, Chen Yingpin, Suqi Huang, Feiyi Qin, and Zhenming Peng. Mask sparse representation based on semantic features for thermal infrared target tracking. *Remote Sensing*, 08 2019. 3

[22] Xin Li, Qiao Liu, Nana Fan, Zhenyu He, and Hongzhi Wang. Hierarchical spatial-aware siamese network for thermal infrared object tracking. *Knowledge-Based Systems*, 12 2018. 3

[23] Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, and Geng Lu. Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In *CVPR*, 2020. 6

[24] Lichao, Zhang, Abel, Gonzalez-Garcia, Joost, van, de, Weijer, Martin, and Danelljan. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 2018. 3

[25] Alan Lukežič, Tomáš Vojíř, Luka Čehovin Zajc, Jiří Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *CVPR*, 2017. 6

[26] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, 2016. 1

[27] J.F.Henriquesand R.Caseiro, P.Martins, and J.Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, March 2015. 6

[28] Jianbo Shi et al. Good features to track. In *CVPR*, 1994. 4, 5

[29] Paul Voigtlaender, Jonathon Luiten, Philip H.S. Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *CVPR*, 2020. 2, 3, 6

[30] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 2

[31] Han Wu, Weiqiang Li, Wanqi Li, and Guizhong Liu. A real-time robust approach for tracking uavs in infrared videos. In *CVPR Workshops*, 2020. 3

[32] X. Yu and Q. Yu. Online structural learning with dense samples and a weighting kernel. *Pattern Recognition Letters*, 2017. 3