A Collaborative Learning Tracking Network for Remote Sensing Videos

Journal:	*IEEE Transactions on Cybernetics*
Manuscript ID	CYB-E-2022-03-0585
Manuscript Type:	Regular Paper
Date Submitted by the Author:	22-Mar-2022
Complete List of Authors:	Li, Xiaotong; Xidian University, School of Artificial Intelligence Jiao, Licheng; Xidian University, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation Zhu, Hao; Xidian University, the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education Liu, Fang; School of Computer Science and Technology; Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University Yang, Shyuan; Xidian university, School of Artificial intelligence, Key Laboratory of Intelligence Perception and Image Understanding of Ministry of Education of China; Xidian university, School of Artificial intelligence , Key Laboratory of Intelligence Perception and Image Understanding of Ministry of Education of China Zhang, Xiangrong; Xidian university, School of Electronic Engineering Wang, Shuang; Xidian University, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation Qu, Rong; University of Nottingham, Computer Science
Keywords:	remote sensing video, object tracking, deep learning, collaborative learning, attention mechanism

A Collaborative Learning Tracking Network for Remote Sensing Videos

Xiaotong Li, Licheng Jiao, *Fellow Member, IEEE*, Hao Zhu, *Member, IEEE*, Fang Liu, *Senior Member, IEEE*, Shuyuan Yang, *Senior Member, IEEE*, Xiangrong Zhang, *Senior Member, IEEE*, Shuang Wang, *Senior Member, IEEE*, and Rong Qu, *Senior Member, IEEE*

***Abstract*—With the increasing accessibility of remote sensing videos, remote sensing tracking is gradually becoming a hot issue. However, accurately detecting and tracking in complex remote sensing scenes is still a challenge. In this paper, we propose a collaborative learning tracking network for remote sensing videos, including a consistent receptive field parallel fusion module (CRFPF), dual-branch spatial-channel co-attention (DSCA) module, and geometric constraint re-track strategy (GCRT). Considering the small-size objects of remote sensing scenes are difficult for general forward networks to extract effective features, we propose a CRFPF-module to establish parallel branches with consistent receptive fields to separately extract from shallow to deep features and then fuse hierarchical features adaptively. Since the objects and their background are difficult to distinguish, the proposed DSCA-module uses the spatial-channel co-attention mechanism to collaboratively learn the relevant information, which enhances the saliency of the objects and regresses to precise bounding boxes. Considering the interference of similar objects, we designed a GCRT-strategy to judge whether there is a false detection through the estimated motion trajectory and then recover the correct object by weakening the feature response of interference. The experimental results and theoretical analysis on multiple data sets demonstrate our proposed method's feasibility and effectiveness. Code and net are available at https://github.com/Dawn5786/CoCRF-TrackNet.**

***Index Terms*—Remote sensing video, object tracking, deep learning, collaborative learning, attention mechanism.**

I. Introduction

In recent years, with the diversification of remote sensing application scenarios, single-frame static remote sensing images can no longer meet the demand for dynamic detection of ground objects. While video satellites can obtain time-series dynamic images of observation areas, which can provide rich information for many applications like traffic condition monitoring, the rapid response of natural disaster, and military security [1]. Object tracking is one of the key technologies in video analysis and understanding applications [2]. Many advanced trackers for natural videos have been proposed [3–5].

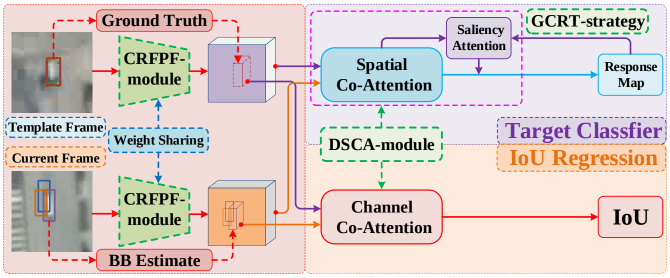


Fig. 1: The overall process framework, including a consistent receptive field parallel fusion module (CRFPF), a dual-branch spatial-channel co-attention module (DSCA), and geometric constraint re-track strategy (GCRT).

The emergence of very high-resolution (VHR) video satellites provides the possibility for remote sensing video tracking. Since remote sensing videos are taken from high altitudes with wide angles, complex scenes and the feature distribution of ground objects are vastly different from those in natural videos. Therefore, accurate real-time remote sensing video tracking remains a particular challenge.

Object tracking usually establishes an appearance model with objects marked in the first frame, then detecting and positioning a designated object in subsequent frames. A general tracking framework usually consists of a search strategy, feature extraction, and observation model. According to the different feature extraction methods, trackers can be mainly divided into traditional methods and deep learning methods [6].

In traditional methods, based on different ways of observation, it can be divided into generative and discriminant models [7]. Generative models usually extract object features to construct an appearance model, such as optical flow [8], Kanade Lucas Tomasi (KLT) [9], Meanshift [10] *etc*. However, the generative model does not make full use of background information and appearance changes. While discriminant models, like tracking learning detection (TLD) [11] and struck [12], usually compare the difference of object and background through a discriminant function. Later, MOSSE [13] first introduces the correlation filtering method to object tracking. On this basis, kernel correlation filter (KCF) [14] introduces cyclic matrix, and fast Fourier transform for real-time online training. These models can take advantage of the relationship between the object and the background simultaneously, but the trained filters limit the universality of the models.

This work was supported in part by the State Key Program of National Natural Science of China under Grant 61836009, in part by the National Natural Science Foundation of China under Grant U1701267, 61871310, 61573267, and in part by the Major Research Plan of the National Natural Science Foundation of China under Grant 91438201 (*Corresponding author: Licheng Jiao*).

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi Province 710071, China (e-mail: lixiaotong@stu.xidian.edu.cn, lchjiao@mail.xidian.edu.cn).

In deep learning methods, trackers use deep features with powerful representation ability instead of manual features [15–17]. Some early works, such as efficient convolution operators (ECO) [18] and hedged deep tracking (HDT) [19], directly use existing pre-trained models to extract deep features. However, due to the non-universality of the pre-trained models, end-to-end learning trackers are proposed which can be trained off-line, such as SiamATL [20], ATOM [21] and Dimp [22], *etc*.

In recent three years, some tracking methods peculiar to remote sensing videos have come into being. Du *et.al.* [23] combine KCF [14] and three-frame difference algorithm to build a strong tracker. Guo *et.al.* [24] design a correlation filter incorporated with a Kalman filter (CFKF) to correct the tracking trajectory of moving targets. Wang *et.al.* [25] design a Gabor filter on CSK [26] to enhance object features. Later, Hu *et.al.* [27] extract deep features with pre-trained deep neural networks. PASIAM [28] uses a shallow siamese network to match object features and predict attention to deal with occlusion. Although the above studies have achieved good performance, the research remote sensing video tracking is still in its infancy. There are still some significant challenges as follows:

1) Since satellite videos are taken from high altitudes, the interesting objects are usually small-size with little sufficient information. However, the present trackers do not sufficiently extract series of hierarchical features for these small objects in particular.

2) Due to the complex atmospheric medium in remote sensing videos (e.g. clouds, fog), objects are usually similar to those of the surrounding background. It makes accurate tracking difficult. However, as far as we know, existing trackers cannot adaptively enhance objects in dynamic backgrounds.

3) When the object to be tracked moves around those similar objects, the tracker is easily disturbed and drifts to the wrong objects. It dramatically affects the success rate of trackers.

Based on the analysis above, we propose a deep collaborative learning tracking network for remote sensing videos. The main contributions can be summarized as follows:

1) For small objects to be detected, we design some parallel multi-resolution branches so that they can have the consistent receiving field on the same level layer, thus extracting the hierarchical features of small objects from shallow to deep layers, and finally, use adaptive weights to fuse them effectively.

2) To strengthen the objects, we use the cross-correlation of objects between the current frame and the template frame to perform collaborative attention learning in spatial and channel , respectively, which can expand the difference between the objects and the background. And an additional attention-loss is designed to enhance the saliency of the objects further.

3) For the tracked results, we design a re-track strategy to judge whether the objects are false tracked through geometric constraint, and then weaken the feature response of the interference objects and finally re-track the actual objects.

The remainder of this paper is organized as follows: in Section II, we introduce three modules of the proposed framework in detail; in Section III, contains the analysis of the experimental results and the effect of each module; finally, Section IV concludes this paper.

II. Methodology

A. Consistent Receptive Field Parallel Fusion Module

In remote sensing videos, the objects are usually small-size, which causes the bottleneck of tracking accuracy. In recent works [29, 30], feature pyramid structures are constructed to meet the challenge of small objects. However, they are usually applied in global recognition on multiple categories of objects at any scale widely. So it lacks specificity in detecting small objects. Based on this, we construct a consistent receptive field parallel fusion module (CRFPF).

The input of this module is an RGB image patch $I_1^0 \in \mathbb{R}^{M_1^0\times M_1^0\times 3}$, which is intercepted by the center of the object.

1) **Construction of Parallel Image Pyramid Input.** Single resolution input is difficult to balance the extraction of deep features and shallow features for extremely small objects. In order to further extract rich hierarchical effective semantic representation information, we construct parallel image pyramid by a multi-resolution sampling way on original image patch I_1^0 and obtain a series of image patches $\{I_0^0,\cdots,I_k^0,\cdots,I_K^0\}$ with corresponding sample rate $\{\alpha_0,\cdots,\alpha_k,\cdots,\alpha_K\}$, where $I_k^0 \in \mathbb{R}^{M_k^0\times\ M_k^0\times 3}$ and M_k^0 represents the side length of I_k^0.

$$M_k^0 = \alpha_k \times M_1^0 \quad \{k\in\mathbb{N}|0\le k\le K\} \tag{1}$$

These image patches correspond to the inputs of $K+1$ parallel convolution branches $\{\mathbb{B}_{I_0},\cdots,\mathbb{B}_{I_k},\cdots,\mathbb{B}_{I_K}\}$. According to the principle of deep neural network, the higher resolution branch tends to extract deep semantic features, while the lower resolution branch tends to extract shallow detailed features.

Among them, $\mathbb{B}_{I_0}$ is a down-sampling branch with $\alpha_0 < 1$ to capture global location feature information from a lower resolution image patch.

So the steps to construct the input process are:

(a) The original image patch I_1^0, enhanced by the red border, is the input of the module.

(b) Then calculate sampling rates $\{\alpha_0,\cdots,\alpha_k,\cdots,\alpha_K\}$.

(c) Finally perform multi-resolution sampling on I_1^0, and obtain $\{I_0^0,\cdots,I_k^0,\cdots,I_K^0\}$ as inputs for parallel branches.

2) **Feature Extraction with Consistent Receptive Field.** For image pyramid as input, after passing through different parallel general convolutional network branches, the actual scene regions corresponding to the receptive fields of the obtained features are not necessarily consistent. So it is not conducive to subsequent feature fusion. In order to resist the misalignment phenomenon during the fusion of non-same branch features, we aim to keep the consistent receptive field on the same level layer for all branches (See the yellow dotted lines in Fig. 2). Note that in a branch, a series of consecutive convolutions form a level layer, their size is fixed. And one branch has L level layers in Fig. 2. In this regard, we think of using dilation convolution [31] which can change receptive field with the same size convolution kernel to align features. Dilation rates of the level layers in branch $\mathbb{B}_{I_k}$ is recorded as $\{r_k^1,\cdots,r_k^l,\cdots,r_k^L\}$.

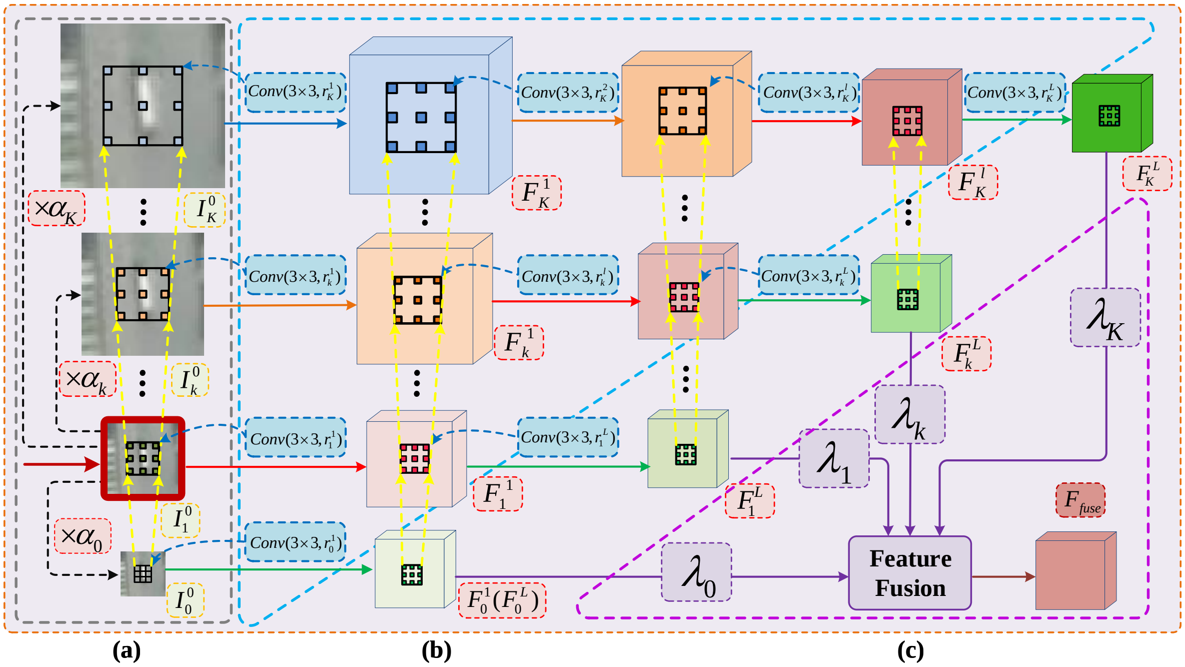


Fig. 2: Flowchart of CRFPF-Module. The input is I_1^0 and the output is fused feature F_{fuse}. 1) Construction of Parallel Multi-Resolution Input is in the grey dashed circle. 2) Feature Extraction with Consistent Receptive Field is in the blue dashed circle. 3) Adaptive Feature Fusion are in the purple dashed circle.

Starting from the branch $\mathbb{B}_{I_1}$, its original RGB image patch $I_1^0 \in \mathbb{R}^{M_1^0 \times M_1^0 \times 3}$ passes through dilation convolution operations to generate the first level layer feature $F_1^1 \in \mathbb{R}^{M_1^1 \times\ M_1^1 \times 3}$ as follows.

$$M_1^1 = \lfloor \frac{M_1^0 + 2 \times padding - r_1^1 \times (Ker - 1) - 1}{stride_1^1} + 1 \rfloor \quad (2)$$

where Ker is the convolution kernel size and r_1^1 is the dilation rate of the first level layer in $\mathbb{B}_{I_1}$.

Among parallel branches, given sampling rate $\alpha_k, (k = \{0, 1, \cdots, K\})$ and dilation rate r_1^1 of branch $\mathbb{B}_{I_1}$, dilation rates r_k^1 of each branch in the first level layer can be calculated as follows.

$$\frac{r_k^1 \times (Ker - 1) + 1}{r_1^1 \times (Ker - 1) + 1} = \alpha_k = \frac{M_k^0}{M_1^0} \quad (3)$$

Therefore, the obtained $\{r_1^1, \cdots, r_K^1\}$ are substituted into Eq. (2) and $\{F_0^1, \cdots, F_K^1\}$ are determined. We design the above rules to ensure the features of the same level layer have a consistent receptive field among different branches. That is they respond to the same real scene range of the original image patches.

For the branch $\mathbb{B}_{I_k}$, with r_k^1 of each branch obtained from the above, the receptive field $\mathfrak{F}_k^i$ of the i-th layer $(i = \{1, \cdots, L\})$ is shown in following recurrence formula as Eq. (4), where the dot represent dot multiplication operations in maths.

$$\mathfrak{F}_k^i = \begin{cases} r_k^i \times (Ker - 1) + 1 & , \quad i = 1 \\ \mathfrak{F}_k^{i-1} + [r_k^i \times (Ker - 1) + 1] \cdot \prod_{p=1}^{i-1} stride_k^p & , \quad i \geq 2 \end{cases} \quad (4)$$

In this process, the current level layer features of the lower resolution branch have the same dimension as the next level layer feature of its adjacent higher resolution branch (*e.g.* F_{k-1}^{l-1} and F_k^l in Fig. 2). So far, features of all branches can be determined. The specific steps are:

(a) First build a standard convolutional network branch B_{I_1} as the feature extraction branch of the initial image I_1^0.

(b) Then obtain parallel branches $\mathbb{B}_{I_k}, (k = \{0, 1, \cdots, K\})$ with given sampling rate $\alpha_k, (k = \{0, 1, \cdots, K\})$ and dilation rate r_1^1 of branch $\mathbb{B}_{I_1}$ as Eq. (2-4).

(c) Finally, $\{I_0^0, \cdots, I_k^0, \cdots, I_K^0\}$ are sent to respective branches hierarchical and features $\{F_0^L, \cdots, F_k^L, \cdots, F_K^L\}$ are extracted.

Based on the above description, the output features $\{F_0^L, \cdots, F_k^L, \cdots, F_K^L\}$ extracted from parallel branches are consistent in dimension.

3) **Adaptive Feature Fusion.**

According to the principle of deep learning, objects of various sizes have different degrees of preference on different level features. So the feature fusion way with equal weights is not particularly suitable. Therefore, we propose an adaptive feature fusion way to obtain the final fused feature F_{fuse} as follows.

$$\begin{aligned} F_{fuse} &= \sum_{k=0}^{K} \lambda_k \cdot F_k^L \\ s.t. \quad \lambda_k &= \frac{\beta_k}{\sum \beta} \end{aligned} \quad (5)$$

where λ_k represents the fusion weight of feature F_k^L, which is normalized by β_k to interval $[0, 1]$. And the dot represent dot multiplication operations.

The proportion β_k is determined by the object size $\sqrt{w \times h}$,

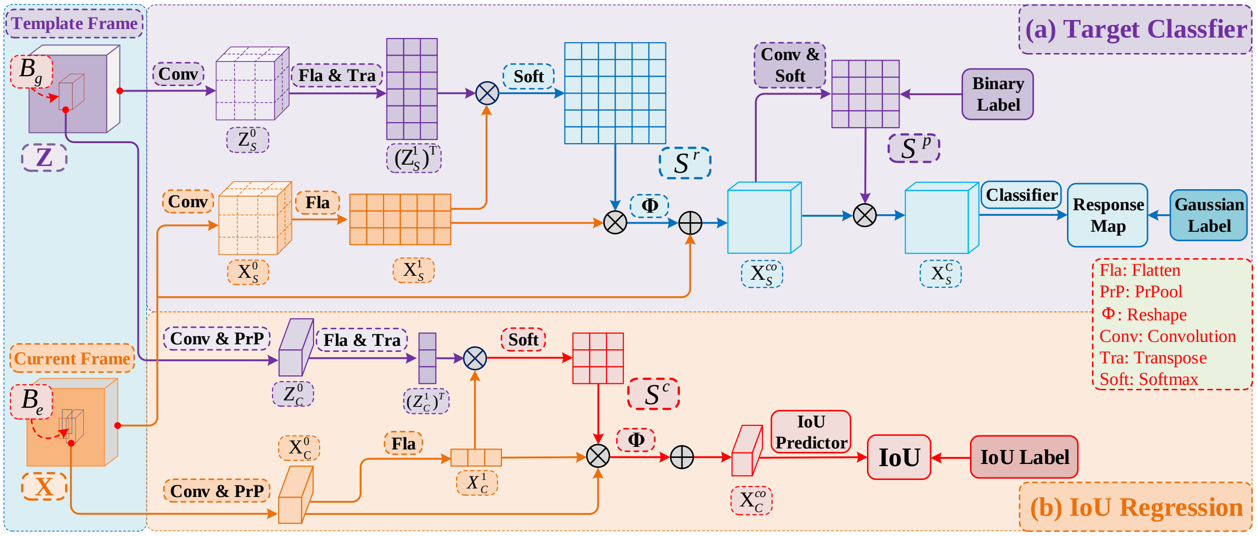


Fig. 3: Flowchart of DSCA-Module. (a) Target Classfier. The inputs are fusion features Z and X of the template frame and the current frame. (b) IoU Regression. The inputs are ground truth region B_g within template frame feature Z and proposal region B_e within current frame feature X, denoted as $Z(B_g)$ and $X(B_e)$.

the image patch size M_1 and the sample rate α_k as follows.

$$\beta_k = \left[\ln(\frac{M_1}{\sqrt{w \times h}})\right]^{-\alpha_k} \tag{6}$$

Given a certain object size, β_k decreases with the increase of α_k, that is, a branch with a larger sample rate has a smaller fusion weight. Besides, given the sample rate of each branch, as the object size increases, the weight gap among the branches decreases and tends to be even. In this way, the smaller the object, the proportion of shallow detail features is larger, conducive to the precise detection of small objects. Function $\ln(\cdot)$ is used to prevent violently jitter of fusion weight caused by object size changes.

In general, for CRFPF-module, multi-resolution parallel branches fully extract the low-level detailed information and high-level semantic information of small objects from shallow to deep layers, consistent receptive fields allow features of different branches in the same level layer corresponding to the same real scene range. And the adaptive fusion way not only contains multiple layers information, but also reduces the deviation caused by feature misalignment.

B. Dual-branch Spatial-Channel Co-Attention Module

To increase the saliency of objects, we propose a DSCA-module composed of a target classifier and an intersection over union (IoU) regression based on collaborative attention (co-attention) mechanism.

1) **Spatial Co-Attention Module.**

To highlight the object region saliency from the spatial structure, we construct a spatial co-attention module on the target classifier. The structure is shown in Fig. 3 (a).

1. Feature Initialization. The inputs, $Z \in \mathbb{R}^{W\times H\times C}$ and $X \in \mathbb{R}^{W\times H\times C}$, are the fused features extracted from CRFPF-module of the template frame and the current frame. They go through several convolution operations and generate initial features $Z_S^0 \in \mathbb{R}^{W\times H\times C}$ and $X_S^0 \in \mathbb{R}^{W\times H\times C}$.

2. Generate Spatial Co-Attention Map. We calculate the spatial co-attention map S^r describing the cross-correlation information between two initial features, Z_S^0 and X_S^0.

Specifically, flatten Z_S^0 and X_S^0 to $Z_S^1 \in \mathbb{R}^{N\times C}$ and $X_S^1 \in \mathbb{R}^{N\times C}$, where $N = W \times H$. We obtain the spatial co-attention map $S^r \in \mathbb{R}^{N\times N}$ as follows.

$$S^r = softmax(X_S^1 \cdot (Z_S^1)^T) \tag{7}$$

We design the spatial co-attention map S^r to measure the similarity of Z_S^1 and X_S^1 in the corresponding spatial position. The larger value in S^r indicates the higher similarity of corresponding positions, vice versa.

3. Feature Modulation. We modulate spatial co-attention map S^r to initial current frame feature X_s^1 based on the current frame feature X, and the cross-correlation feature $X_S^{co} \in \mathbb{R}^{W\times H\times C}$ is calculated as follows.

$$X_S^{co} = X + \Phi(S^r \cdot X_s^1) \tag{8}$$

where $\Phi(\cdot)$ represents an operation to reshape input back to the initial dimension as X. Therefore, according to the degree of cross-correlation between the template frame and the current frame, we enhance the object region signal and enlarge the difference in the feature distribution between the object and its background.

4. Saliency Attention-Loss. Considering that the feature information of small objects is very minimal, if they are directly sent to the classifier, the feedback loss will not be enough to improve the saliency of the object. Therefore, we design an additional attention-loss to further focus on the neighborhood of objects, thereby further expanding the gap between the object and the background.

For this attention-loss, the saliency mask $S^p \in \mathbb{R}^{W\times H\times 1}$ is generated as follows.

$$S^p = softmax(conv(X_S^{co})) \tag{9}$$

where $conv(\cdot)$ represents several 1×1 convolution operations.

The additional attention-loss function is as follows.

$$L_s(S^p, G) = -\sum(1-\eta)glog(s^p) + \eta(1-g)log(1-s^p) \tag{10}$$

where G is a binary label obtained according to the ground truth boxes. For G, the value of the object location is set to

1, while the background is set to 0. η presents the percentage of positive values in the binary label template to prevent an imbalance between positive and negative samples. $g \in G$ and $s^p \in S^p$. In the tracking process, we multiply the modulation feature X_S^{co} with salient mask S^p, getting final attention feature X_S^C for the target classifier.

Subsequent target classifier consists of several fully convolutional layers. The loss for the classifier is the least square function, as shown in Eq. (11).

$$L_{cls}(X_S^C, Y) = \sum (f(x_s^c; \omega) - y)^2 + \mu \|\omega\|^2 \tag{11}$$

here, Y is the Gaussian sampling label centered at the object, $y \in Y$ and $x_s^c \in X_S^C$. μ is the amount of regularization on ω. The overall loss is as follows,

$$L_{spatial} = \xi L_s + (1 - \xi) L_{cls} \tag{12}$$

where ξ is in $[0, 1]$ to balance the two losses.

2) **Channel Co-attention Module.**

In IoU predictor, each channel of feature has a different contribution to the accurate prediction. To increase the proportion of beneficial channels under the guidance of precise ground truth, we propose a channel co-attention module as Fig. 3 (b).

1. Feature Initialization. The inputs are ground truth region B_g within the template frame feature Z, and the proposal region B_e within the current frame feature X, denoted as $Z(B_g)$ and $X(B_e)$. Both them are fed through convolution operations, and then do precise region of interest pooling operations(PrPool) [32] as Eq. (13) to get continuous down-sampling of the same dimension, getting initial features $Z_C^0 \in \mathbb{R}^{J \times J \times C}$ and $X_C^0 \in \mathbb{R}^{J \times J \times C}$.

$$PrPool(B(x_1, y_1, x_2, y_2)) = \frac{\int_{y1}^{y2} \int_{x1}^{x2} f(x, y) dx dy}{(x_2 - x_1) \times (y_2 - y_1)} \tag{13}$$

where B represents the above B_g or B_e, (x_1, y_1) and (x_2, y_2) represent the coordinates of the upper left and lower right corners of B. $f(x, y)$ represents the element value of the feature map at coordinates (x, y).

2. Generate Channel Co-Attention Map. We flatten Z_C^0 and X_C^0 to $Z_C^1 \in \mathbb{R}^{J^2 \times C}$ and $X_C^1 \in \mathbb{R}^{J^2 \times C}$ along the channel axis and obtain channel co-attention map $S^c \in \mathbb{R}^{C \times C}$ as follows.

$$S^c = softmax((Z_C^1)^T \cdot X_C^1) \tag{14}$$

The S^c obtained in this way reflects the degree of cross-correlation between channels from Z_C^0 and X_C^0, respectively. We design S^c to measure the degree of similarity between channels. Larger values indicate higher similarity, vice versa. According to the cross-correlation score, the channel that is highly similar to the channel of the ground truth object is enhanced.

3. Channel Modulation. We generate channel modulated features $X_C^{co} \in \mathbb{R}^{J \times J \times C}$ with S^c as follows.

$$X_C^{co} = X_C^0 + \Phi(X_C^1 \cdot S^c) \tag{15}$$

where operation $\Phi(\cdot)$ reshape input back to dimension as X_C^0. This not only adaptively adjusts the channel weight, but also completely ensures the original feature information.

4. IoU Estimate. Given X_C^{co}, the IoU value is predicted by several fully connected layers. The IoU regression loss computed is as follows:

$$L_{IoU} = -ln(\frac{B_g \cap B_e}{B_g \cup B_e}) \tag{16}$$

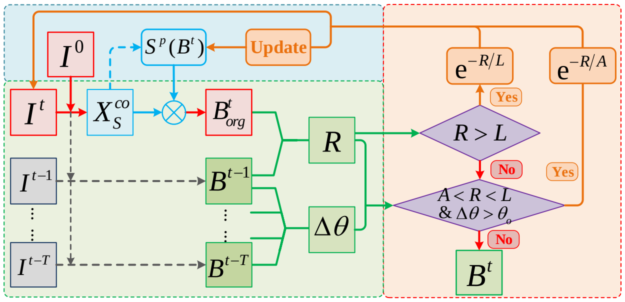


Fig. 4: Flowchart of GCRT strategy of the t-th frame. I_0 is the original input of the template frame, $I_{t-T} \sim I_t$ is the original input of T frames.

Finally, the overall DSCA-model is complete. In this module, we introduce the cross-correlation characteristics between the template frame and the current frame, and make use of the cross-correlation information. In this way, we respectively enhance the saliency of the object region in the spatial domain and prefer high-quality features intelligently in the channel domain. Note that the target classification branch and the IoU regression branch focus on different domains and cannot be interchanged. Primarily, we also design a saliency attention-loss to further enlarge the difference in feature distribution between objects and similar backgrounds.

C. Geometric Constraint Re-Track Strategy

In remote sensing videos, objects are small-size and have little texture information, so the indiscernible appearance of objects leads to the difficulty of retrieving objects once lost. Therefore, we propose a re-track strategy based on geometric constraints to reduce false detections, as shown in Fig. 4.

Given the previous T frames results $\{B^{t-1}, \cdots, B^{t-T}\}$, we judge whether the estimate result $B_{org}^t(x_c^t, y_c^t, w^t, h^t)$ of the current t-th frame needs to be re-track, where (x_c^t, y_c^t) is the center of the box and w^t, h^t represent the width and height. For the box of the t-th frame, we update its corresponding mask S^p response by a attenuation factor φ as follows.

$$S^p(B^t) = S^p(B_{org}^t) \cdot \varphi \tag{17}$$

We discuss φ for re-track in three situations as:

$$\varphi = \begin{cases} e^{-R/L} & , \quad R > L \\ e^{-R/A} & , \quad A < R \le L \quad \& \quad \triangle\theta > \theta_o \\ 1 & , \quad else \end{cases} \tag{18}$$

where R means the Euclidean distance between (x_c^t, y_c^t) and (x_c^{t-1}, y_c^{t-1}). L and A are the diagonal length and the shorter side length of the box, respectively.

$$L = \sqrt{(w^t)^2 + (h^t)^2} \tag{19}$$

$$A = \min(w^t, h^t) \tag{20}$$

1) **R > L.**

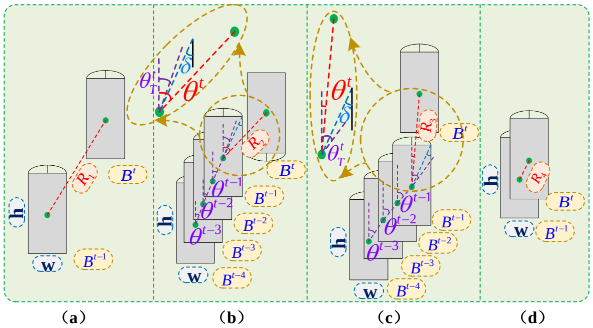


Fig. 5: Display for several possible situations. (a) $R_1 > L$. (b) $A < R_2 \leq L, \theta^t - \theta^t_T$. It shows the situation when similar targets are moving towards each other in close distance; (c) $A < R_3 \leq L, \theta^t_T - \theta^t$. It shows the situation when similar targets are moving in the same direction. (d) It shows the normal situation that $R_4 < A$.

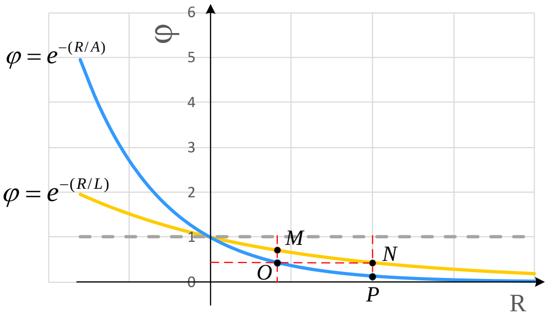


Fig. 6: Attenuation function curve comparison. When $R > 0$, the value range of the function is in the interval [0, 1]. As R increases, the smaller the attenuation factor φ is, the more severe the signal attenuation is. In order to avoid ineffective suppression of interferers at different distances, $e^{-R/L}$ is used when $R > L$, and $e^{-R/A}$ is used when $A < R \leq L$.

In remote sensing videos, since the tracked object is usually a low-speed moving one, the Euclidean distance the object moves between two adjacent frames is usually less than the length of the object itself, *i.e.*, L in Eq. (19).

Therefore, when $R > L$, we consider the estimated result of the t-th frame to be unreliable. So we take $\varphi = e^{-R/L}$ for re-track. It attenuates the unreliable detection object according to the distance. As R is larger, φ is closer to 0, and more severe attenuation. In this way, the unreliable detection objects are weakened so that the real objects can be re-tracked.

2) $\mathbf{A < R \leq L}$.

In this situation, we consider that objects may appear in two positional relationships as shown in Fig. 5 (b)(c). So we further use the angle relationship to make the determination.

We take the average angle change θ^t_T of the previous T frames as a benchmark to predict the current angle change:

$$\theta^t_T = \frac{1}{T}\sum_{i=1}^{T} \arctan \frac{x_c^{t-i} - x_c^{t-(i+1)}}{y_c^{t-i} - y_c^{t-(i+1)}} \tag{21}$$

where x_c^{t-i}, y_c^{t-i} are corresponding center of previous boxes. And the current movement direction is:

$$\theta^t = \arctan \frac{x_c^t - x_c^{t-1}}{y_c^t - y_c^{t-1}} \tag{22}$$

Then we calculate the angle change $\triangle\theta$ of the current frame:

$$\triangle\theta = \mid \theta^t - \theta^t_T \mid \tag{23}$$

We use the angle θ_o between the diagonal and the long side of the object bounding box B^t as the threshold of the angle change range, as follows:

$$\theta_o = \arctan \frac{\min(w^t, h^t)}{\max(w^t, h^t)} \tag{24}$$

Since small objects in our remote sensing scenes are moving at a low speed, $\triangle\theta$ should be an acute less than θ_o.

If $A < R \leq L$ and $\triangle\theta > \theta_o$ are satisfied at the same time, we start the re-track procedure with the attenuation factor as $\varphi = e^{-R/A}$. Compared with $e^{-R/L}$ in the the case of $R > L$, $\varphi = e^{-R/A}$ also exponentially decreases as R increases. The difference is that $\varphi = e^{-R/A}$ decays faster, which can effectively suppress interference even near the object. If the above two situations do not occur, it is considered that there is no false detection, and φ is unchanged as 1. The reason we use two exponential functions with different coefficients is to strike a balance between sensitively capturing false tracking results and avoiding allergy alarms. As shown in Fig. 6, suppose we use a uniform decay function $e^{-R/L}$ indiscriminately, then when $R \leq L$, the value of the φ is relatively high (e.g. $\varphi_M > \varphi_O = \varphi_N$). Therefore, we replace the attenuation function in the interval $A < R \leq L$ with $e^{-R/A}$, so that the nearby interferers can also be effectively attenuated without affecting the magnitude of the long-distance attenuation. In this way, the interference signals at different distances can be effectively attenuated during re-tracking, and the allergy alarm caused by the jitter of bounding boxes can be avoided.

With the above GCRT strategy, we use geometric constraints to weaken the response of the mask S^p in unreliable detection object region, thereby bringing out the correct real object during re-tracking.

III. Experimental Results and Discussion

A. Data Description

1) **IPIU Date Set.**

The data set is acquired over the San Diego Military Port, USA, 2017 by Jilin-1 HD Dynamic Video Satellite with 10 fps. The actual ground resolution is 0.91 m/pixel. The various vehicle objects in the scene are to be tracked. 95% of object sizes are from 5×8 pixels to 10×15 pixels. The scene contains bridges, roads, trees, buildings, and many similar vehicle objects. Less effective target information and complex scenes cause great difficulties in object tracking.

2) **RSSRAI Data Set.**

This data set comes from the remote sensing video target tracking track competition in *2019 Remote Sensing Image Sparse Representation and Intelligent Analysis Competition* [33]. The total size of a frame is from 220×223 pixels to 1348×1348 pixels. These video sequences are shot with a resolution of 1.13 m/pixel. The object sizes are from 4×9 pixels to 18×19 pixels. The object signal strength is weak, so it is difficult to distinguish it from the background.

3) **UAV123* Data Set.**

The public UAV123 data set [34] contains a series of video sequences from an aerial drone viewpoint. In all 123 sequences, we take vehicles as objects and select sequences with BC (Background Clutter) and LR (Low Resolution) attributes, forming the UAV123* data set to conduct experiments. This data set sequence is characterized by a wide range of viewing angles, which leads to an unstable direction of objects.

B. Experimental setup

The overall experiments are conducted on a workstation with Intel Xeon(R) CPU E5-2650 v4 @ 2.20GHz and 4 NVIDIA TITAN Xp GPU. The proposed tracker is implemented on PyTorch deep learning platform.

1) **Backbone Network Learning.**

We use CRFPF-module proposed in Section II-A as our backbone network to fusion and extract features. Considering the balance between computing power and time cost, we take $K = 3$ in Eq. (1) in practice, that means a total of four parallel branches $\{\mathbb{B}_{I_0}, \mathbb{B}_{I_1}, \mathbb{B}_{I_2}, \mathbb{B}_{I_3}\}$. We add a classifier head consisting of a 1×1 convolution layer and a 4×4 convolution layer after the backbone network [21]. We take $\mu = e^{-4}$ in Eq. (11). In order to adapt the backbone network to the remote sensing data set, we use GOT-10K [35] natural data set and 25% IPIU (denoted as IPIU^T) as the training data set. For the pair of selected frames, we crop out the image patches that with several times (5 times in our experiments) the size of the object and centered on the bounding box. Then we resize them to 72×72 as inputs into the backbone with sharing weight. In each epoch, 60% samples are randomly selected from GOT-10K and the remaining 40% from IPIU train set. In order to avoid the situation that it is difficult to converge for multi-branch training at the same time, we adopt an alternate training strategy. When training a specific branch, we set the loss weight of the rest branches to 0. Each branch trains 80 epochs with 2×10^{-4} learning rate.

2) **Target Classifier Learning.**

The Spatial Co-attention Module is trained online to adapt to the foreground and background discrimination of the current test sequence. During training this process, we keep the above backbone weight frozen and only update the weight of the dual-branch Spatial Co-attention Module. For each test sequence, we use the first template frame of the object's ground truth to generate a series of similar but slightly different enhanced frames. In experiments, we generate 15 enhanced samples, then choose any two enhanced samples to form a pair as the input. The weights of attention loss in Eq. (10) and classifier loss in Eq. (11) are ξ and $1-\xi$ respectively, see C(2) in this section for details.

3) **IoU Regression Learning.**

We use image pairs with bounding box annotations to train the entire IoU prediction network. During training this process, we keep the above backbone weight frozen and only update the weight of the dual-branch Channel Co-attention Module. We only use the above IPIU^T as the source of image pairs. We sample image pairs from each sequence with a maximum interval of 30 frames. Then we set displacement similar to [21] to achieve the purpose of expanding the training data. During the training process, each batch includes 26 image pairs, the learning rate is 1×10^{-3}, and a total of 60 epochs are trained.

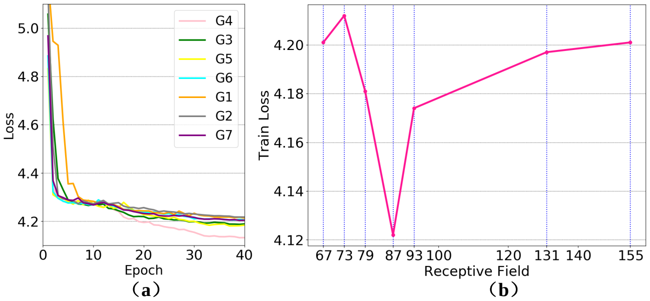


Fig. 7: Comparison of different dilation rate settings. (a) Train total loss of different dilation rate. (b) Convergence loss of series receptive field.

TABLE I: Comparison of average convergence loss with different dilation rates.

	Level1	Level2	Level3	Absolute Receptive Field	Convergrnce Loss
G_1	1	1	1	67	4.201
G_2	2	1	1	73	4.212
G_3	3	1	1	79	4.181
G_4	3	2	1	87	4.122
G_5	4	2	1	83	4.174
G_6	5	4	3	131	4.197
G_7	5	5	5	155	4.201

C. Hyperparameter Analysis

1) **Dilation rate r_k.**

In Section II-A, the dilation rate r_k is used to adjust consistent receptive field among parallel branches. Different dilation rate settings correspond to different receptive fields, which has a great impact on the distinguish ability of features for small objects. According to Eq. (3), given the dilation rate r_k of certain branch I_k, the dilation rate of rest branches can be derived according to the consistent receptive field principle. Therefore, we select a single I_2 branch with a moderate number of network layers for experiments. In the experiments, we record the group form of dilation rates as

TABLE II: The structure of each branch in CRFPF-Module.

Branch	Input Shape	Output Shape	Layer	r_*^*	*stride*	*padding*
$\mathbb{B}_{I_0}$	(36,36,3)	(18,18,64)	Conv[7×7,64]	1	2	1
	(18,18,64)	(18,18,64)	Conv$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	1	1	1
$\mathbb{B}_{I_1}$	(72,72,3)	(36,36,64)	Conv[7×7,64]	2	2	3
	(36,36,64)	(18,18,64)	Pool[3×3]	–	2	1
	(18,18,64)	(18,18,64)	Conv$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	1	1	1
$\mathbb{B}_{I_2}$	(144,144,3)	(72,72,64)	Conv[7×7,64]	3	2	3
	(72,72,64)	(36,36,64)	Pool[3×3]	–	2	1
	(36,36,64)	(36,36,128)	Conv$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 2$	2	1	1
	(36,36,128)	(18,18,64)	Conv$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	1	2	1
$\mathbb{B}_{I_3}$	(288,288,3)	(144,144,64)	Conv[7×7,64]	7	2	3
	(144,144,64)	(72,72,64)	Pool[3×3]	–	2	1
	(72,72,64)	(72,72,64)	Conv$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	4	1	1
	(72,72,64)	(36,36,128)	Conv$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 1$	2	2	1
	(36,36,128)	(18,18,64)	Conv$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	1	2	1

TABLE III: Performance comparison with different K.

Values of K	Target Classier Loss	IoU Loss	Total Loss	Time
K=1	0.656	0.183	4.121	1h 27min
K=2	0.648	0.172	4.059	2h 13min
K=3	**0.639**	**0.172**	**3.997**	**3h 3min**
K=4	0.669	0.182	4.119	5h 51min
K=5	0.679	0.192	4.275	8h 32min

TABLE IV: Comparison of multiple spatial loss weight ξ.

ξ	Convergence Loss	Convergence Epoch
0.001	4.248	48/50
0.01	4.085	46/50
0.1	3.987	45/50
0.3	3.990	45/50
0.5	3.992	46/50
0.7	4.273	45/50
0.9	4.281	45/50
0.99	4.288	42/50
0.999	4.366	41/50

$G_i = \{r_i^1, r_i^2, r_i^3, i = 1 \sim 7\}$, where r_i^1, r_i^2, r_i^3 represent the dilation rate of Level1, Level2, and Level3 in branch B_2. We use the 1×1 area of the last features corresponding to the real area of the original input (denoted as Absolute Receptive Field) as a unified comparison index. From G_1 to G_7, the Absolute Receptive Field gradually increases as shown in Table I. The rest parameters are kept consistent for fair.

The results are shown in Fig. 7. As the receptive field increases, convergence loss first increased slightly, then decreased rapidly to valley at 87, and finally slowly recovered. So we select G_4 with minimum convergence loss. The specific structure of each branch is shown in Table II.

2) **Number of branches K.**

In Section II-A, K is an important value worth choosing in feature extraction.

Considering the objects in our datasets are uniformly from 5×8 pixels to 10×15 pixels in size, we design each branch as Table II and conduct an overall test. The experimental results are shown in Table III.

At the beginning, the accuracy of the network increase with increasing K value. When $K = 3$, the accuracy of the network has risen to a stable level. We consider that the network at this time has been able to sufficiently extract the deep features of small objects. When $K > 3$, the accuracy of the network is no longer improved, but the computational complexity is still multiplying. Considering the balance between the parameter amount and accuracy, we take $K = 3$. In fact, the value of K can be adjusted according to the size of the object itself, the magnitude of the training dataset, and timeliness requirements in specific practical application.

3) **Spatial loss weight ξ.**

In Section II-B, ξ in Eq. (12) is the weight of losses in the total spatial loss $L_{spatial}$. Its value affects the convergence of the network. Therefore, in order to make the value of ξ cover a range as wide as possible, we set ξ exponentially in a wide range of intervals [0.001, 0.999]. During the experiment, the rest of the parameter settings are the same. The experimental results are shown in Table IV.

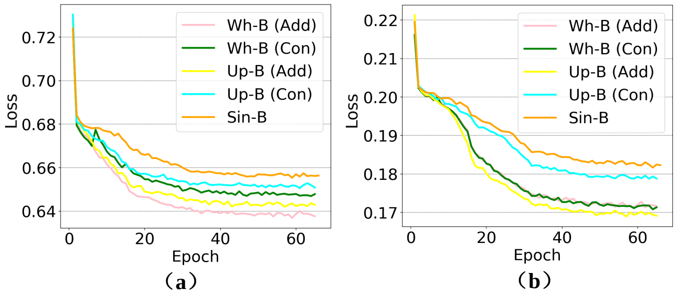


Fig. 8: Comparison of different combinations in CRFPF-module. (a) Target classifier loss of different branch combinations. (b) IoU regression loss of different branch combinations.

TABLE V: Performance comparison of different combinations in CRFPF-module.

Branches Combination	Target Classier Loss	IoU Loss	Total Loss	Time
Sin-B (K=1)	0.656	0.183	4.121	1h 27min
Up-B (*Con*)	0.651	0.179	4.088	7h 58min
Up-B (*Add*)	0.642	0.169	4.019	3h 3min
Wh-B (*Con*)(K=3)	0.646	0.171	4.047	8h 25min
Wh-B (*Add*)(K=3)	**0.639**	**0.172**	**3.997**	3h 27min

As we can see, the overall loss first decreases and then increases with the increases of ξ from 0.001 and is lowest when $\xi = 0.1$. The convergence speed increases slowly as ξ increases while of the same magnitude. When the loss percentage (on the same magnitude) is dominated by L_{cls} and assisted by L_s, the model has a more vital ability to distinguish between target and background. It also confirms the effectiveness of the introduced attention loss. Therefore, we set ξ to 0.1 finally.

D. Module Comparison

We conduct comparative experiments individually on each module proposed in Section III. To ensure the experiments' fairness, only the settings of the module to be compared are different while the rest keep consistent.

1) **CRFPF-Module.**

In the experiment, Sin-B, Up-B, and Wh-B are backbone networks with different branch combinations. Sin-B represents a single branch, Up-B represents non-downsampling branch fusion, Wh-B represents whole branches (upsampling and downsampling) fusion. The letters in brackets indicate different fusion methods: *Add* stands for additive fusion, *Con* stands for concatenate fusion. The specific structure of each branch is shown in Table II. The experimental results are shown in the Table V and Fig. 8.

Sin-B *Vs.* Up-B:

As shown in Fig. 8, in both the target classifier and the IoU regression training, the training loss of Up-B converges to a lower value than that of Sin-B. (Note that, Sin-B uses a single best-performing branch $\mathbb{B}_{I_2}$). As a result, the features of different levels of non-downsampling branches are fused, which is more comprehensive than a single branch's features. For small objects, single-resolution feature expression

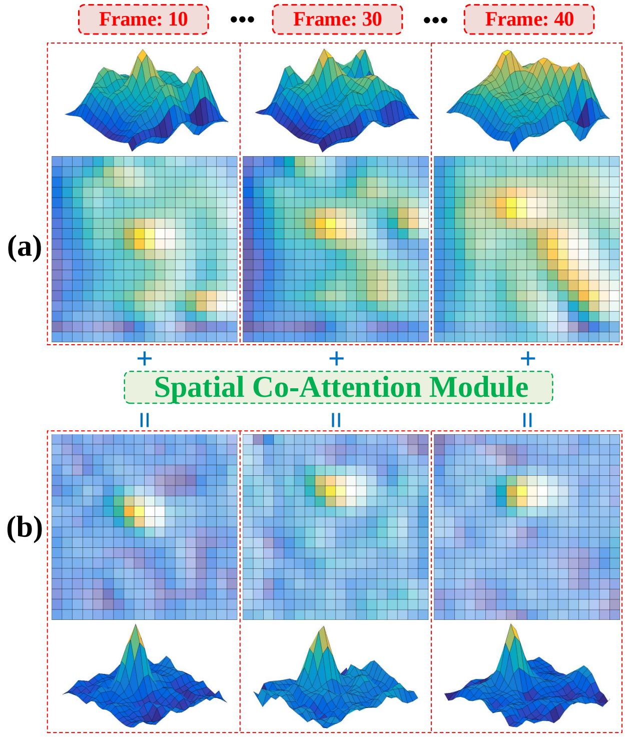


Fig. 9: Qualitative comparison of response maps on spatial co-attention module. (a) Response maps of the original algorithm without co-attention. (b) Our response maps.

is usually insufficient, while the hierarchical features of multi-resolution can be used to supplement, which improves the overall performance. However, the training time of the network increased at the same time.

Up-B *Vs.* Wh-B:

Wh-B adds down-sampling branch $\mathbb{B}_{I_0}$ based on Up-B. It can be seen from the experimental results, for the target classifier, whether the fusion way is an additive or concatenate operation, the training convergence loss of Wh-B is lower than that of Up-B. It is verified that the introduced shallow global features have positively affect the target classifier. However, since the IoU regression uses the precise position feature with PrPool, the addition of global information in $\mathbb{B}_{I_0}$ does not show apparent advantages.

Add Fusion *Vs. Con* Fusion:

For the target classifier, since the features are consistent in the spatial structure, additive fusion can strengthen the target information and performs better than concatenate fusion. Moreover, in terms of training time, the training efficiency of additive fusion is doubled higher than that of concatenating fusion.

In summary, we choose the optimal structure of Wh-B *Add*.

2) **DSCA-Module.**

We conduct comparative experiments on the spatial co-attention target classifier and the channel co-attention IoU regression to verify the effect of our DSCA-module.

Spatial Co-Attention:

Fig. 9 shows the changes in the feature map's response before and after adding the spatial co-attention module. By learning the correlation of the object between the current frame and the template frame, we strengthen the regions that are strongly correlated with the template object. Moreover because the attention loss is specially trained, the object region is further highlighted, while the response of the similar surrounding background is weakened. As shown in Fig. 9, compared with the original algorithm without co-attention, the peak of our response maps are concentrated in the object region, and its surrounding interference is suppressed.

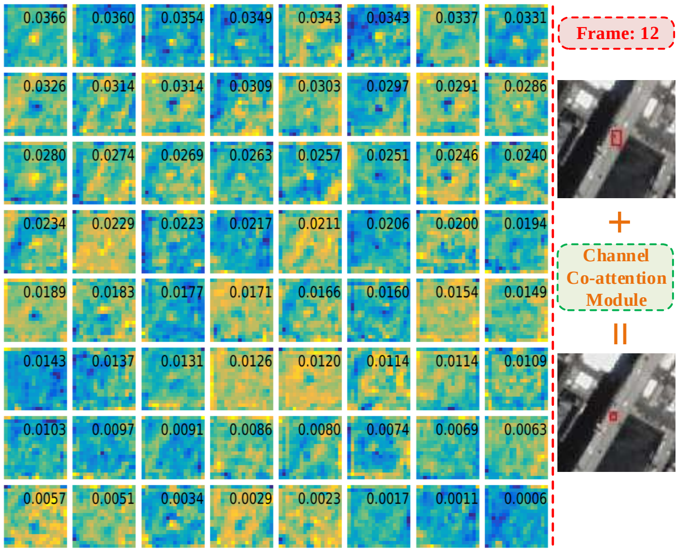


Fig. 10: Visualization of channel features and corresponding weight by our channel co-attention module. Take the 12-th frame as an example to show the comparison of the results before and after adding the Channel Co-attention Module.

Channel Co-Attention:

Fig. 10 shows the comparison of the tracking results before and after adding the channel co-attention module. According to the cross-correlation between the features of the current frame and the template frame, for object features, the channels that are highly similar to that of the ground truth object are enhanced. In this way, channels containing more object information help the IoU's accurate prediction, and it also has a better understanding of the object boundary. For example, Fig. 10 shows all 64 channel features of the 12-th frame in a particular sequence and also shows the corresponding weights assigned by the channel co-attention. We can see that most channels with more significant object information have higher weights. The final results in Fig. 10 show that the tracking result with the channel co-attention module is tighter and more precise.

3) **GCRT-Strategy.**

Fig. 11 shows the effect of the GCRT-strategy with a fragment of a sequence. Our strategy uses geometric relations to determine whether the object is lost, and re-tracks to the correct object by weakening the response of the interfering obnect in the saliency mask. As shown in Fig. 11 . When encountering an interfering object facing each other, our strategy keeps the target from being lost.

E. Ablation Studies and Algorithm comparison

In this section, we conduct comparative experiments with several state-of-art methods and self-ablation experiments. For the sake of experimental fairness, we adopt standard indicators Precision and Success in single object tracking area according to literature [36] to evaluate the performance of trackers.

TABLE VI: Ablation Study and Performance Comparison Results of Three Data Sets.

	IPIU			RSSRAI			UAV123*		
	Precision	Success	FPS	Precision	Success	FPS	Precision	Success	FPS
DCF [14]	0.516	0.156	428.782	0.261	0.084	338.482	0.407	0.334	455.448
SAMF [37]	0.546	0.179	20.995	0.360	0.133	16.963	0.403	0.263	24.813
DLSSVM [38]	0.750	0.414	100.319	0.536	0.340	83.592	0.206	0.039	104.636
LCT [39]	0.544	0.040	35.795	0.348	0.159	33.697	0.394	0.274	43.012
ECO [18]	0.856	0.484	15.167	0.781	0.431	13.136	0.529	0.282	16.760
HDT [19]	0.266	0.008	46.817	0.168	0.031	31.672	0.153	0.013	51.864
DIMP [22]	0.468	0.215	18.361	0.674	0.441	14.733	0.661	0.612	22.612
DIMP-A	0.521	0.241	22.541	0.691	0.511	17.495	0.698	**0.648**	25.280
DIMP-B	0.505	0.224	17.230	0.675	0.479	14.933	0.681	0.612	25.957
DIMP-C	0.517	0.238	15.282	0.670	0.500	13.476	0.696	0.639	20.072
ATOM [21]	0.794	0.395	20.731	0.729	0.437	16.954	0.440	0.285	23.696
base-A	0.828	0.419	25.060	0.783	0.449	19.827	0.653	0.511	27.293
base-AB	0.877	0.454	16.103	0.784	0.494	14.932	0.661	0.599	19.692
base-ABC (Ours)	**0.954**	**0.554**	14.051	**0.816**	**0.566**	12.516	**0.729**	0.621	15.671

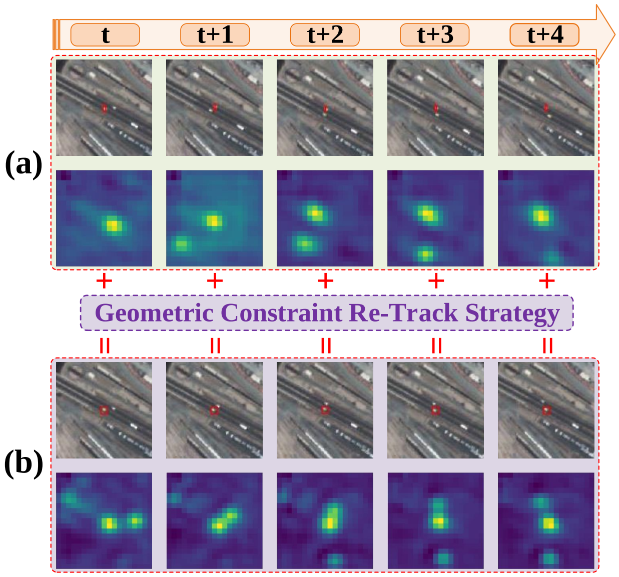


Fig. 11: The response map is generated with the detected bounding box as the center. (a) Tracking result and response map of sample sequence without re-tracking strategy. (b) Tracking result and response map of sample sequence with re-tracking strategy.

We compare our tracker framework with some state-of-art trackers, including DCF [14], SAMF [37], DLSSVM [38], LCT [39], ECO [18], HDT [19], DIMP [22], and baseline ATOM [21]. DCF is a representative of the discriminant kernel correlation filtering method. SAMF use feature pyramid correlation filter to perform multi-scale tracking. DLSSVM and LCT are algorithms that combine detection and tracking. ECO and HDT algorithm combine the powerful deep learning presentation. For the above algorithms, we adapt their official codes without changing the network structure. DIMP and ATOM integrate deep network and filtering classification into an end-to-end trainable tracker. Their frameworks are similar to ours so we use the consistent initialization of the same settings.

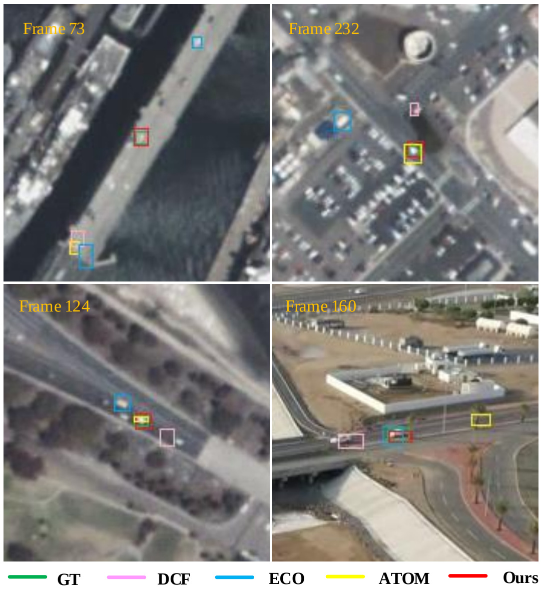


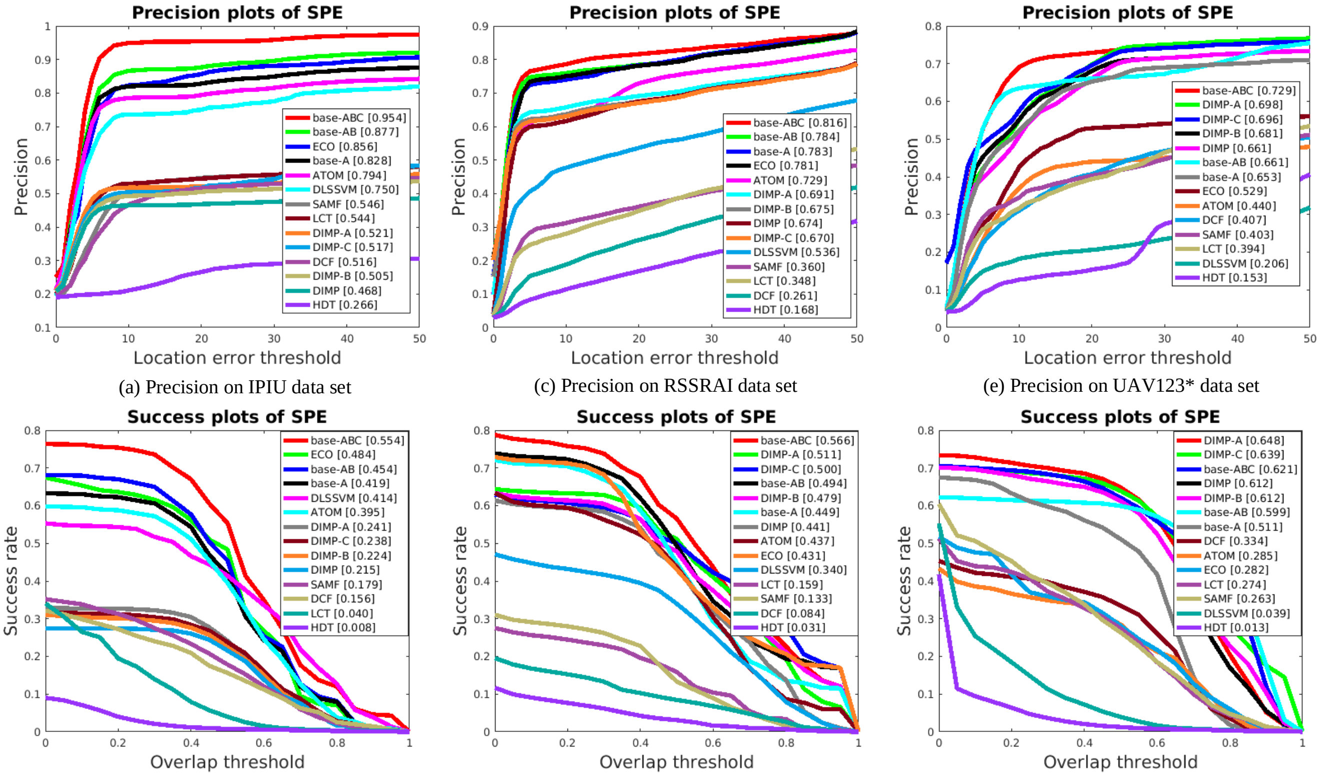
Fig. 13: Display of tracking results of DCF, ECO, ATOM and our algorithms

We carry out a quantitative ablation study taking ATOM as a baseline algorithm to verify each proposal component of our framework, including base-A, base-AB, and base-ABC. A represents CRFPF-module, B represents DSCA-module, and C represents GCRT-strategy. The quantitative results are shown in Table VI.

1) **Experiment Analysis on IPIU Data Set**.

DCF is a correlation filtering algorithm that is known for its speed advantage. As shown in Table VI, the number of tracking frames per second is up to hundreds. However, shallow manual design features, *i.e.* directional gradient histogram (HOG) and color names (CN), have low robustness to complex remote sensing scenes. Therefore, it is susceptible to interference from multiplicative background noise.

SAMF is a multi-scale method that focues on the situation where the size of the object changes drastically. However, the object sizes in our remote sensing videos are small and the change is not apparent, so the tracking Success is not



(a) Precision on IPIU data set

(c) Precision on RSSRAI data set

(e) Precision on UAV123* data set

(b) Success on IPIU data set

(d) Success on RSSRAI data set

(f) Success on UAV123* data set

Fig. 12: Comparison of different algorithms on multiple algorithms.

significantly improved.

DLSSVM uses a traditional structured SVM kernel to determine the object and LCT algorithm also use a re-detector composed of SVM to re-detect hard negative samples. However, it is difficult to achieve stable performance on complex scenes due to weak adaptive learning ability.

ECO tracker that combines superficial appearance information and deep semantic information is significantly better than HDT tracker with only convolutional features, despite the speed price. However, they two directly use the pre-trained model on ImageNet data set, which limits the trackers' learning ability and scalability from remote sensing videos.

DIMP and ATOM allow trainable learning of remote sensing data characteristics. However, too deep forward network reduces the feature discrimination of small objects. When similar objects appear around, the tracker is challenging to distinguish. And it is almost difficult to retrieve the original object once a mistake occurs. Although DIMP has online update ability, low-quality features become a burden after the tracking fails.

Compared with baseline ATOM, base-A merges deep and shallow features in parallel to represent small size objects. It can effectively extract the deep features of small objects and fuse them with the shallow features to improve the robustness of features while ensuring the integrity of spatial detail information. Experiment results in Table VI show 1.52 and 1.42 increases in Precise and Success, respectively.

Compared with base-A, base-AB improves the saliency of the object in the spatial domain, weakens the response of the similar surrounding background, and enhances the dominant feature in the channel domain. Therefore, its Precise outperforms ECO. However, due to its algorithm is not enough to distinguish the moving interference object effectively, the result of Success does not reach the ideal level.

Our complete tracking framework, base-ABC, further takes the geometric constraint re-detection strategy, achieves a gain of 13.6 in precise score and 15.9 in success score in comparison with baseline. The GCRT-strategy uses time context information to constrain accidental error. The hard-to-recover plight in remote sensing scenarios is alleviated.

In addition, the three proposed modules can also be effectively transferred to another baseline algorithm DIMP, denoted as DIMP-A, DIMP-B, and DIMP-C, respectively. They all bring about performance improvements in Success and Precision. This demonstrates the extensibility of the module.

For time complexity, as shown in Table VII, the baseline ATOM is $O(13.67E+09)$, while the proposed network is $O(18.91E+09)$, which is 1.38 times that of ATOM. For space complexity, the parameters of the baseline ATOM are about $T(11.16E+06)$, while the overall parameters of the proposed network is about $T(13.92E+06)$. (i.e. CRFPF-module with $K=3$ is $T(9.27E+06)$), so they are within an order of magnitude. Compared with ATOM, our methods tracks about 6 fewer frames per second. The reduction of each part is as follows: the CRFPF-module increases by 5 frames; in DSCA-module, the cross-correlation calculation in the collaborative attention mechanism reduces by about 9 frames; the GCRT-strategy reduces by about 2 frames. Therfore, our proposed methed keeps the overall speed within an order of magnitude of ATOM while ensuring Success and Precision.

2) **Experiment Analysis on RSSRAI Data Set**.

The experimental results on RSSRAI data set are shown in the Fig. 12 (c)(d). It can be seen that tracking methods fused

TABLE VII: Model Complexity Analysis.

		Time Complexity		Space Complexity	
		Backbone	Total	Backbone	Total
ATOM	Parameters	O(9.95E+09)	O(13.67E+09)	T(5.57E+06)	T(11.16E+06)
	Memory	37.05GB	50.94GB	21.23MB	42.58MB
base-ABC (Ours)	Parameters	O(15.16E+09)	O(18.91E+09)	T(9.27E+06)	T(13.92E+06)
	Memory	56.49GB	70.44GB	35.35MB	53.11MB

with deep learning like ATOM, DIMP is better than traditional correlation filter, SVM frameworks. This is mainly due to the background changes in some tracking sequences. For example, when a vehicle drives to the transition between a dark asphalt road and a bright bridge, the background suddenly changes. In this way, the DIMP tracker with online learning capabilities can learn the changes in time, and the tracking Success is higher than the offline training tracker ATOM.

In addition, since the resolution of the RSSRAI data set is 1.13m/pixel, the object blur is more serious. When re-tracking strategy is used alone, the accuracy of bounding boxes obtained during regression is affected, although the target object can be recovered. So Precision (i.e. 0.670) slightly decreases by 0.004 compared with that of DIMP (i.e. 0.674) even if its Precision improves.

According to the visualized results of tracking, it is found that when the vehicle turns at an intersection, the vehicle direction and shape are quite different from the initial state, especially who with a long body. Therefore, the performance of ECO in RSSRAI is not as good as that in IPIU. Moreover, base-AB tracker we proposed uses the object information of the template frame to strengthen the similar target area of the frame to be tested, shields the interference of drastic changes in the background to a certain extent, and improves the Success by 5 percentage points. This factor limits the increase in Precision.

From the point of view of speed, although image size of RSSRAI is larger than that of IPIU and our tracking framework still has the same level on FPS compared with deep learning trackers, such as DIMP.

3) **Experiment Analysis on UAV123* Data Set.**

The experimental results are shown in the Fig. 12 (e)(f). Since it belongs to a short-distance shooting by drones, images contains more spatial detail information, and the shape and scale of the target change more diversely, the ATOM and DIMP trackers with IoU-Net branch perform pixel-level regression on the rough box during tracking process. In this case, the performance is better than the ECO algorithm estimated with multiple scales. Thereby, due to online updated classifier module, DIMP perform better than the baseline ATOM on UAV123* data set. So Success of DIMP and its extension algorithm are higher than these of ATOM. DCF captures object information through hand-designed features, so it has good rotation invariance and robustness to the change of view angle. While the object captured in UAV123* data set has angle variability, so the results of DCF in UAV123* is better than that of the previous two data sets. In the contrary, DLSSVM is sensitive to appearance and shooting angle, and its performance is not good enough as CF trackers.

In the ablation experiment, the object scene captured by UAV123* is more complex, and it is more vulnerable to the interference of the surrounding moving targets. Our base-A can effectively extract the hierarchical robust features from the shallow to the deep, and the overall base-ABC with GCRT-strategy can also effectively eliminate the surrounding moving targets. The Precision and Success are improved by about 0.28 and 0.33 respectively on UAV123*.

IV. Conclusion

In this paper, we propose a collaborative learning network applied to remote sensing video tracking. Experiments have verified that CRFPF-module provides a idea to extract effective hierarchical features especially from small objects; DSCA-module collaboratively learns the object commonality between the template frame and the current frame, and uses spatial channel correlation to highlight the weak target signal; GCRT-strategy provides a re-tracking strategy for remote sensing of complex large scenes, reducing the interference of similar moving objects. However, the training process of the DSCA-module relies on a large number of annotated tracking frames, which limits the flexibility of the strategy in practical application; in addition, the efficiency of our algorithm is not particularly high. In the future, we will further study on how to lighten the structure of the backbone network, reduce the dependence on the amount of training data, save computational costs in the future.

Acknowledgment

The authors would like to thank Dr. Wenhua Zhang at the Xidian University of China for her kind assistances about comparative experiments.

References

[1] S. K. Patel and A. Mishra, "Moving object tracking techniques: A critical review," *Indian Journal of Computer Science and Engineering*, vol. 4, no. 2, pp. 95–102, 2013.

[2] L. Jiao, R. Zhang, F. Liu, S. Yang, and X. Tang, "New generation deep learning for video object detection: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–21, 2021.

[3] W. Zhang, L. Jiao, Y. Li, and J. Liu, "Sparse learning-based correlation filter for robust tracking," *IEEE Transactions on Image Processing*, vol. 30, pp. 878–891, 2020.

[4] X. Wang, Z. Hou, W. Yu, Z. Jin, Y. Zha, and X. Qin, "Online scale adaptive visual tracking based on multilayer convolutional features," *IEEE transactions on cybernetics*, vol. 49, no. 1, pp. 146–158, 2017.

[5] T. Feng, L. Jiao, H. Zhu, and L. Sun, "A novel object re-track framework for 3d point clouds," in *Proceedings of*

the 28th ACM International Conference on Multimedia, 2020, pp. 3118–3126.

[6] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 4, pp. 1–47, 2020.

[7] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V. D. Hengel, "A survey of appearance models in visual object tracking," *ACM transactions on Intelligent Systems and Technology*, vol. 4, no. 4, pp. 1–48, 2013.

[8] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[9] J. Shi *et al.*, "Good features to track," in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 1994, pp. 593–600.

[10] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[11] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1409–1422, 2011.

[12] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. Torr, "Struck: Structured output tracking with kernels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2096–2109, 2015.

[13] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2544–2550.

[14] J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[15] H. Zhu, L. Jiao, W. Ma, F. Liu, and W. Zhao, "A novel neural network for remote sensing image matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2853–2865, 2019.

[16] M. Liu, L. Jiao, X. Liu, L. Li, F. Liu, and S. Yang, "C-CNN: Contourlet convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[17] L. Jiao, R. Shang, F. Liu, and W. Zhang, *Brain and Nature-inspired Learning, Computation and Recognition*. Elsevier, 2020.

[18] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6638–6646.

[19] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M. H. Yang, "Hedged deep tracking," in *Computer Vision and Pattern Recognition*, 2016.

[20] B. Huang, T. Xu, Z. Shen, S. Jiang, B. Zhao, and Z. Bian, "SiamATL: Online update of siamese tracking network via attentional transfer learning," *IEEE Transactions on Cybernetics*, 2021.

[21] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4660–4669.

[22] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6182–6191.

[23] B. Du, Y. Sun, S. Cai, C. Wu, and Q. Du, "Object tracking in satellite videos by fusing the kernel correlation filter and the three-frame-difference algorithm," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 168–172, 2017.

[24] Y. Guo, D. Yang, and Z. Chen, "Object tracking on satellite videos: A correlation filter-based tracking method with trajectory correction by kalman filter," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3538–3551, 2019.

[25] Y. Wang, T. Wang, G. Zhang, Q. Cheng, and J. Q. Wu, "Small target tracking in satellite videos using background compensation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–12, 2020.

[26] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *European conference on computer vision*. Springer, 2012, pp. 702–715.

[27] Z. Hu, D. Yang, K. Zhang, and Z. Chen, "Object tracking in satellite videos based on convolutional regression network with appearance and motion features," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 783–793, 2020.

[28] J. Shao, B. Du, C. Wu, and Y. Pingkun, "Pasiam: Predicting attention inspired siamese network, for space-borne satellite video tracking," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1504–1509.

[29] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3578–3587.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proceedings of International Conference on Learning Representations*, 2016, pp. 2881–2890.

[32] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 784–799.

[33] RSSRAI, "http://rscup.bjxintong.com.cn/," 2019.

[34] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," *Far East Journal of Mathematical Sciences*, vol. 2, no. 2, pp. 445–461, 2016.

[35] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[36] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[37] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *European conference on computer vision*. Springer, 2014, pp. 254–265.

[38] J. Ning, J. Yang, S. Jiang, L. Zhang, and M.-H. Yang, "Object tracking via dual linear structured svm and explicit feature map," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4266–4274.

[39] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5388–5396.

Xiaotong Li received the B.S. degree in electronic information engineering from the Harbin Engineering University, Harbin, China, in 2017. She is currently pursuing the Ph.D. degree with Xidian University, Xian, China. Her main research interests include deep learning, computer vision and remote sensing image analysis and understanding.

Licheng Jiao (Fellow, IEEE) received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xian Jiaotong University, Xian, China, in 1984 and 1990, respectively. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, Xian, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. His research interests include image processing, natural computation, machine learning, and intelligent information processing. Dr. Jiao is a Foreign Member of the Academia Europaea and the Russian Academy of Natural Sciences; a fellow of IET, CAAI, CIE, CCF, and CAA; a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, and an Expert of the Academic Degrees Committee of the State Council, the Chairman of the Awards and Recognition Committee, and the Vice Board Chairperson of the Chinese Association of Artificial Intelligence.

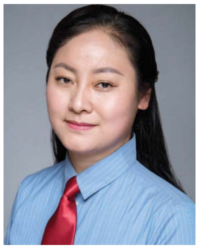
Hao Zhu received the B.S. degree in physics and photoelectricity engineering and the Ph.D. degree in Circuits and Systems from Xidian University, Xian, China, in 2013 and 2019, respectively.

He is currently a Lecturer with the school of Artificial Intelligence, the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University of China. His current research interests include deep learning, remote sensing image interpretation, and evolutionary computation.

Fang Liu (Senior Member, IEEE) received the B.S. degree in computer science and technology from Xian Jiaotong University, Xian, China, in 1984, and the M.S. degree in computer science and technology from Xidian University, Xian, in 1995. She is currently a Professor with Xidian University. She has authored or coauthored five books and over 80 papers. Her research interests include image perception and pattern recognition, machine learning, evolutionary computation. Prof. Liu won the second prize of the National Natural Science Award in 2013.

Shuyuan Yang (Senior Member, IEEE) received the B.A. degree in electrical engineering and the M.S. and Ph.D. degrees in circuit and system from Xidian University, Xian, China, in 2000, 2003, and 2005, respectively. She has been a Professor of artificial intelligence with Xidian University. Her research interests include machine learning and image processing.

Xiangrong Zhang (Senior Member, IEEE) received the B.S. and M.S. degrees in computer application technology from the School of Computer Science, Xidian University, Xian, China, in 1999 and 2003, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the School of Electronic Engineering, Xidian University, in 2006. From January 2015 to March 2016, she was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University. Her research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.



Shuang Wang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in circuits and systems from Xidian University, Xian, China, in 2000, 2003, and 2007, respectively. She is a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. Her research interests include sparse representation, image processing, synthetic aperture radar (SAR) automatic target recognition, remote sensing image captioning, and polarimetric SAR data analysis and interpretation.

Rong Qu (Senior Member, IEEE) is currently an Associate Professor with the School of Computer Science, University of Nottingham, Nottingham, U.K. Her research interests include automated algorithm design for real-world optimization based on hybrid computational optimization algorithms in operational research and artificial intelligence. Dr. Qu is also the Vice-Chair of the Evolutionary Computation Task Committee of the IEEE Computational Intelligence Society. She is also an Associate Editor of the IEEE Computational Intelligence Magazine, the IEEE TEVC, the Journal of the Operational Research Society, and the Computer Science (PeerJ). She has guest edited special issues on the automated design of search algorithms and machine learning at the IEEE TAMI and IEEE Computational Intelligence Magazine.

A Collaborative Learning Tracking Network for Remote Sensing Videos

IEEE Transactions on Cybernetics

March 22, 2022

Xiaotong Li, *Licheng Jiao**, *Fellow Member, IEEE*, Hao Zhu, *Member, IEEE*, Fang Liu, *Senior Member, IEEE*, Shuyuan Yang, *Senior Member, IEEE*, Xiangrong Zhang, *Senior Member, IEEE*, Shuang Wang, *Senior Member, IEEE*, and Rong Qu, *Senior Member, IEEE*
e-mail:lixiaotong@stu.xidian.edu.cn,lchjiao@mail.xidian.edu.cn

The Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education,
International Research Center for Intelligent Perception and Computation Joint International
Research Laboratory of Intelligent Perception and Computation
School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi Province 710071, China

Contents

The Response to Editor-in-Chief

Dear Miss Li:

Based on the referee reports and the Associate Editor's recommendation, I regret to inform you that your paper cannot be accepted and have to be rejected. We hope that the attached reviews are useful to you in your research. If the reviewers have attached a PDF or text file review, you will need to retrieve it from your Author Center. Therefore, please check the Author Center for more information when appropriate.

In view of the reviews and recommendation of the Associate Editor, we would like to encourage you to revise the manuscript thoroughly to address the issues raised by the reviewers/Associate Editor and resubmit the revised version as a new submission when ready along with the paper ID as well as your specific responses to the reviewers.

Thank you for submitting to the IEEE Transactions on Cybernetics.

Sincerely,

Peng Shi, FIEEE, FIET, FIEAust

Editor-in-Chief, IEEE Transactions on Cybernetics

Authors' Response:

Dear Editor,

Please find our revised manuscript entitle "A Collaborative Learning Tracking Network for Remote Sensing Videos" (Manuscript No.CYB-E-2021-03-0587). We sincerely appreciate the reviewers' comments and feel encouraged by their positive feedback. For reviewers' concerns and their opinions on improving the manuscript, we have made point-to-point replies and corrections to the revised manuscript, and marked red in the revised manuscript.

Moreover, we also make the following improvements in the revised manuscript:

The point-to-point respond to the reviewer's comments are listed as following.

Should you have any questions, please contact us without hesitation. We are looking forward to your response.

Sincerely

Xiaotong Li, $LichengJiao^*$, Hao Zhu, Fang Liu, Shuyuan Yang, Xiangrong Zhang, Shuang Wang and Rong Qu.

P.O. Box 224, No.2 South Taibai Road, Xi'an 710071, P.R. China.

Tel: +86 15771753228.

Email: lixiaotong@stu.xidian.edu.cn; lchjiao@mail.xidian.edu.cn

The Response to Associate Editor

Associate Editor Comments for Authors:

Comments to the Author:

Three review reports have been received. All reviewers have some technical concerns about this manuscript. The authors are encourged to address all of their comments when preparing a revised version.

Authors' Response:

Dear Associate Editor:

Thank you for your letter and for the reviewers' comments concerning our manuscript entitled "A Collaborative Learning Tracking Network for Remote Sensing Videos" (Manuscript No.CYB-E-2021-03-0587). Those technical comments are all valuable and very helpful for revising and improving our paper, as well as the important guiding significance to our researches. We have studied comments carefully and have made correction which we hope meet with approval. Revised portion are marked in red in the paper. The main corrections in the paper and the responds to the reviewers' comments are as following:

1 Response to Reviewer 1

Comments:

This manuscript proposes a collaborative learning tracking network for remote sensing video. It mainly includes three parts, the authors first propose a consistent receptive field parallel fusion module (CRFPF) to deal with the challenge of small goals, then propose a dual-branch spatial-channel co-attention (DSCA) module, to uses the spatial-channel co-attention mechanism to collaboratively learn the relevant information of the Target Classifier and IoU Regression, finally use geometric constraint re-track strategy (GCRT) to trace back the results of previous frames to re-evaluate and correct the tracking results. In my opinion, the manuscript is well written and clearly understandable. The most attractive feature of this manuscript is its novelty, and the obtained results are convincing.

However, I have some questions and suggestions as follows.

Reviewer 1 Comment 1

For the challenge of small object, the current mainstream methods usually use feature pyramid for feature fusion among different feature levels. Why do the authors use image pyramid in input? What are the unique differences and advantages compared with the feature pyramid? In addition, what is the significance of keeping the consistent receptive field on the same level layer for all branches?

Response

First of all, we thank the Reviewer very much for the valuable comments about our manuscript. We are very sorry for not providing a clear and complete description of the image pyramid and consistent receptive field So we elaborate on this part and re-written it according to the Reviewer's suggestion.

1) The reason why we use image pyramid as input is:

In remote sensing videos, for a single resolution input image, the convolutional features of small objects will be lost as the neural network deepens, making it difficult to extract. Yet this is precisely the key for accurate identification and tracking. Therefore, we augment the input source image into a multi-resolution image pyramid structure to further enrich their semantic feature representation of small objects.

As we know, remote sensing images are usually very large in area, and the width and height may reach several thousand pixels. In comparison, our targets are almost extremely small objects with side lengths in the range of [5, 10] pixels. If a deep neural network is used to directly extract image features, the feature information will be drastically reduced as the network deepens or even loses. The small object features obtained in this way lack deep semantic information, so they have poor robustness in recognition. In order to retain the shallow detail features and obtain the deep semantic features simultaneously, we augment the original image to multi-resolution spaces and construct a series of image pyramid inputs. In this way, multi-level hierarchical features can be extracted in parallel, which enriches the feature space of small objects. Therefore, we utilize the image pyramid structure to address the challenge of small object feature extraction.

2) The difference between our image pyramid and the usual feature pyramid is:

Feature pyramid takes a single source resolution image as input, which extracts multi-level features with different scales from different layers of a neural network. While image pyramid augments the source image into a series of multi-resolution images as input, and it extracts hierarchical features simultaneously with the same scale from different layers of multiple parallel branches.

In some image processing tasks such as segmentation or detection, feature pyramid is a common structure for extracting multi-scale features. These tasks are usually global recognition, requiring a good detector to recognize multiple categories objects at any scale widely. So feature pyramid structure must take targets of various scales into account simultaneously. The specific implementation method is to extract features of different scales from different layers of the network. As the network goes from shallow to deep, the feature scale also goes from large to small, forming a structure similar to a pyramid. This achieves the purpose of taking into account both large and small goals.

However, in our remote sensing object tracking task, the target to be tracked and detected is mainly a single designated extremely small object. Such as a moving vehicle, whose side length scale is about $5-10$ pixels. Thus, the feature pyramid is not very suitable for our task scenario. In order to focus on the hierarchical features of small objects, we extend the original resolution image to multi-resolutions and form an image pyramid structure. In this way, we focus on extracting small object features and avoiding redundancy.

3) The advantage of image pyramid structure is:

Image pyramid structure can further extract rich hierarchical effective representation information, including semantic features for even extremely small objects. The original image is expanded to a multi-resolution image pyramid series, which expands multi-scale space for feature extraction of small objects. This designed image pyramid structure can extract local shallow detail features and deep semantic features at the same time, making subsequent tracking of remote sensing small objects more accurate and refined. Therefore, we use the image pyramid structure as input for feature extraction of small objects in specific remote sensing scenarios.

4) The significance of keeping the consistent receptive field on the same level layer for all branches:

For input image pyramid, after passing through different parallel general convolutional network branches, the actual scene regions corresponding to the receptive fields of the obtained features are not necessarily consistent. Even though the features' dimensions are of the same size (which can be directly fused in form), there is confusion and misalignment during feature fusion. In order to resist the misalignment phenomenon during the fusion of non-same branch features, we must maintain the consistent receptive field on the same level for all branches. Therefore, we perform precise calculations (Eq. (2)- Eq. (4) in the manuscript) to ensure that the actual scene regions of receptive fields corresponding to features are consistent. That is, consistent receptive fields allow features of different branches in the same level corresponding to the uniform real scene regions.

Considering the Reviewer's suggestion, we have carried out a more detailed and precise description in Section II A) in our manuscript. Special thanks for your valuable comments again!

Reviewer 1 Comment 2

Why do the authors use spatial co-attention in the Target Classifier branch, while use channel co-attention in the IoU Regression branch? Can these two different attention modules be exchanged between branches? If not, please give reasons.

Response

We thank the reviewer for your meaningful comments. We are very sorry for our unclear expression in Dual-branch Spatial-Channel Co-Attention Module.

1) The purpose of using the spatial co-attention mechanism in the Target Classifier branch is to enhance the saliency of small target areas with weak signals. In remote sensing images, with the existence of atmospheric media, light scattering, and other multiplicative noise, the boundary between the target signal and the surrounding background environment is blurred and difficult to distinguish. The task of the target classification branch is to identify the approximate area of the target in the spatial search area. Therefore, it is reasonable to use the spatial co-attention mechanism to find regions of interest by enhancing the signal of the weak target area.

2) The purpose of using the channel co-attention mechanism in the IoU Regression branch is to obtain an accurate regression box from the candidate area. These candidate regions are multi-channel features obtained by sampling around the region of interest. And then, they are pooled to a uniform size in the target classification branch. It is worth noting that these multiple channels of the features have different contributions to the accurate prediction of IoU: Some channels in the feature make the prediction of the regression box more accurate, and some channels may make the prediction box result worse. In order to adjust the proportion of the channel under the guidance of the accurate ground-truth box in the template frame, we design the channel co-attention module rather than the spatial co-attention module to help regress to precise target bounding boxes.

3) The two different attention modules between branches can not be exchanged.

Because the tasks of the two branches are different, the role of the co-attention mechanism of the two branches is different too. The Target Classification branch is to identify the approximate area of the target in the search area. It is necessary to find the approximate area of the target in the spatial domain according to the highest point of the probabilistic response. Therefore, the signal at the target needs to be enhanced spatially. In the IoU Regression branch, a set of candidate regions are sampled from rough target area. These candidate regions have very little and similar spatial information. While features in the channel domain are diverse, each channel has a different contribution to the accurate prediction of IoU. So we design a channel co-attention mechanism to guide the prediction of the target bounding box with the precise ground truth features of the template frame. In other words, we utilize channel co-attention to allocate the channel ratio of the current frame feature to maximize the use of accurate template ground truth information.

We revise our manuscript according to your suggestions in Section II B). Thank you for your comments to make our presentation clearer again.

Reviewer 1 Comment 3

In Figure 4, why are the settings of attenuation functions designed to be different? What's the difference between the two? Which is more relaxed? Why is it so designed?

Response

Thank you very much for your valuable suggestions. We are very sorry for not writing clearly in this regard.

1) The reason we design different settings of attenuation functions designed are as follows:

We take the moving vehicle as a representative (i.e., one kind of object in our remote sensing videos) to analyze the motion characteristics of the tracking target. During the movement of the vehicle, there are various complex positional relationships with surrounding similar interfering vehicles. A single function is not conducive to accurately describing the scene in all different situations, such as distant interfering vehicles, nearby interfering vehicles moving toward or away from each other, and so on. Therefore, we analyze the motion state of the vehicles and design different decay functions under different situations.

2) The design principle of the attenuation function is as follows:

Specifically, compared with the broad global scene size in remote sensing videos, the displacement of objects such as vehicles between adjacent frames is minimal. So they are regarded as low-speed moving objects. The Euclidean distance of displacement (denoted R, calculated based on the center point of the target box) between two adjacent frames is usually less than the length of the object itself. Therefore, we can use the diagonal length of the object bounding box itself as a reference. When the displacement between two adjacent frames of the object is greater than the length of the diagonal line, it indicates that the tracking result does not conform to the motion criterion of the object, and mistracking is considered. We denote the diagonal length of the object bounding box as L, as shown in Eq. (1) (Eq. (19) in manuscript).

$$L = \sqrt{(w^t)^2 + (h^t)^2} \tag{1}$$

When $R > L$, we enable the re-tracking procedure to attenuate the signal of the interferer area. We design an attenuation function to obtain the attenuation factor φ. Since the attenuation factor φ needs to be multiplied by corresponding mask $S^p(B^t_{org})$ to achieve interference signal suppression. It is observed that the value of the functional form $e^{-R/L}$ is in the interval $[0, 1]$. So it is equivalent to attenuating the unreliable detection object according to the distance. The larger R is, the closer φ is to 0, and the attenuation is more severe. The change law of the function is in line with the attenuation needs of the actual situation. So we take $\varphi = e^{-R/L}$ as the attenuation function for re-tracking. In this way, unreliable distant detection objects are weakened so that real objects can be tracked again.

When $R \leq L$ we continue to consider the case: If only $R \leq L$ is satisfied, it is not enough to judge whether the tracking result is the normal movement of the object, the common regression deviation, or the wrong tracking to the very close interference object. So the conditions for judging whether it is mistracking should be stricter. Therefore, the re-tracking conditions are not only limited by distance but also by angle. Through geometric mapping, we find that if the tracker mistakenly follows the similar interfering objects close to the target object, a displacement of $(0.5w_{target} + 0.5w_{interfer})$ will be generated at least. (Here, w represents the shortest side length of the object bounding box.) We approximately think that the size of the interferer is similar to the size of the target object. So we set A as a critical threshold as Eq. (2) (Eq. (20) in manuscript). Therefore, when distance relationship $A < R \leq L$ and angular relationship $\Delta\theta > \theta_0$ (details in Section II C)) are satisfied at the same time, we start the re-track procedure with the attenuation factor as $\varphi = e^{-R/A}$.

$$A = \min(w^t, h^t) \tag{2}$$

where w^t, h^t represent the width and height of the target bounding box.

When $R < A$, we consider the displacement as the object's normal motion or regression deviation. Therefore no re-tracking process is triggered, then $\varphi = 1$.

To sum up the above, the overall expression of the attenuation factor φ can be expressed as Eq. (3) (Eq. (18) in manuscript):

$$\varphi = \begin{cases} e^{-R/L} & , \quad R > L \\ e^{-R/A} & , \quad A < R \leq L \quad \& \quad \triangle\theta > \theta_o \\ 1 & , \quad else \end{cases} \tag{3}$$

Similar with $\varphi = e^{-R/L}$ in the previous case of $R > L$, the attenuation function $\varphi = e^{-R/A}$ is also in the interval $[0, 1]$ and exponentially decreases as R increases. However, they are slightly different in the design of the exponential coefficients.

3) The difference between the two attenuation function:

We make an intuitive comparison through the curve of the function. As shown in Fig. 1, the yellow curve shows the function $\varphi = e^{-R/L}$, and the blue curve shows the function $\varphi = e^{-R/A}$. Correspondingly, M and N are points on function $e^{-R/L}$; O and P are points on function $e^{-R/A}$. It can easily observe from the figure that both two functions are monotonically decreasing functions. When $R > 0$, the value range of the function is in the interval [0, 1]. As R increases, the smaller the attenuation factor φ is, the more severe the signal attenuation is. The rule conforms to the needs of the actual scene: when the displacement exceeds the threshold, the farther the distance is, the less likely it is to be the target.

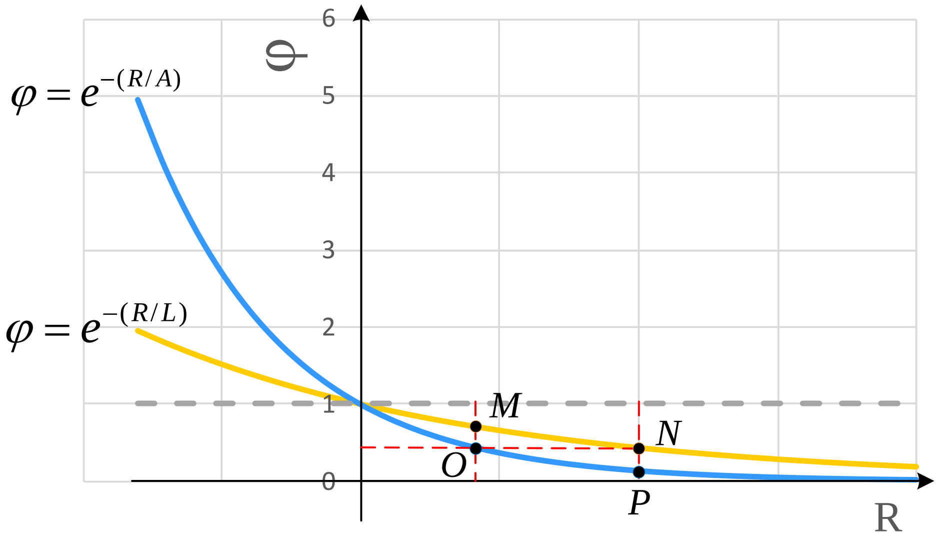


Figure 1: Attenuation function curve comparison.

The reason we use two exponential functions with different coefficients is to strike a balance between sensitively capturing false tracking results and avoiding allergy alarms. Suppose we use a uniform attenuation function $e^{-R/L}$ indiscriminately, then when $R \leq L$, the value of the attenuation factor φ is relatively high (e.g. $\varphi_M > \varphi_O = \varphi_N$). Therefore, we replace the attenuation function in the interval $A < R \leq L$ with $e^{-R/A}$, so that the nearby interferers can also be effectively attenuated without affecting the magnitude of the long-distance attenuation. In this way, the interference signals at different distances can be effectively attenuated during re-tracking, and the allergy alarm caused by the jitter of bounding boxes can be avoided.

4) Through the above description, the constraint in the case of $R > L$ is more relaxed.

In the case of $R > L$, only distance relation constraint needs to be satisfied to start the re-track procedure. While in the case of $R \leq L$, both the distance relation constraint $A < R \leq L$ and the angular relation constraint $\Delta\theta > \theta_0$ need to be satisfied to execute the re-tracking procedure. So the constraint for $R > L$ is more relaxed.

In general, the attenuation function we designed is in line with the actual situation. And we have re-written this part according to the Reviewer's suggestion in Section II C) in our manuscript. Thanks again for your wise opinion.

Reviewer 1 Comment 4

Is the angle range in the geometric constraint limited to a small acute angle In Figure 5? Is there an obtuse angle?

Response

Special thanks to you for your good comments. We apologize for the unclear description of the angle range in the manuscript.

In remote sensing videos, since the object to be tracked is usually a low-speed moving one, the Euclidean distance the object moves between two adjacent frames is generally less than the length of the object as the description above (i.e., $R \leq L$). Based on this rule of motion, after performing the geometric motion analysis shown in Fig. 2 (Fig. 5 in manuscript), we find that the diagonal angle θ_o of the vehicle body is the critical threshold when two different vehicles meet. Because the bounding box of the vehicle is a rectangle, it can be known from the geometric knowledge that the diagonal angle of the rectangle is in the range of [0, 90], so the critical angle in the angular determination should be acute.

Of course, this is the geometric rule of remote sensing objects under the condition of slow motion. If in other scenes, such as natural video or video with a low frame rate, the object may no longer meet the condition of slow motion, then it is possible that the critical angle value range is expanded to an obtuse angle. Therefore, when using the Geometric Constraint Re-Track Strategy (GCRT-Strategy) in other scenes, it is necessary to design the critical angle rule that meets the situation according to the motion law in the specific scene. We have already supplemented this point in Section II C).

Reviewer 1 Comment 5

In the verification experiment, whether the initialization of several comparison algorithms is consistent or not, note that the fairness of the experimental comparison algorithm is very important, please make further elaboration.

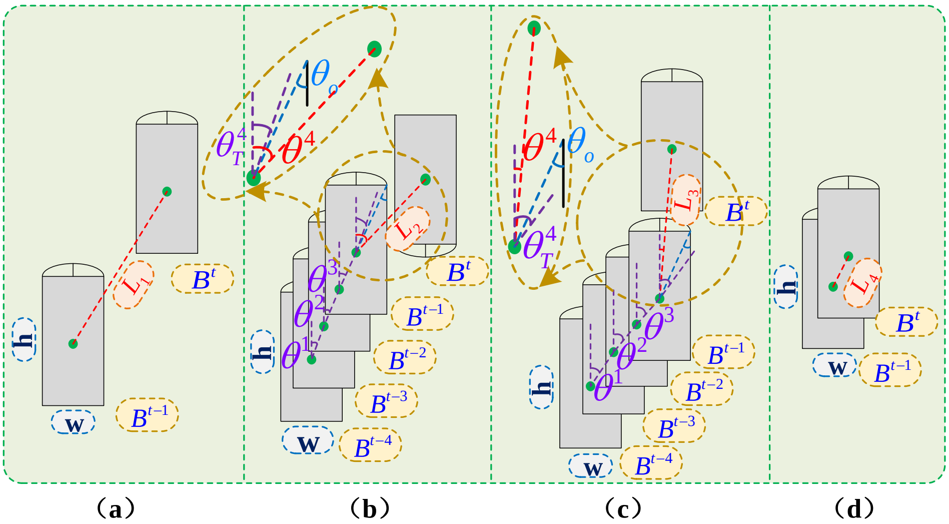


Figure 2: Display for several possible situations. (a) $R_1 > L$. (b) $A < R_2 < L, \theta^t - \theta_T^t$. It shows the situation when similar targets are moving towards each other in close distance; (c) $A < R_3 < L, \theta_T^t - \theta^t$. It shows the situation when similar targets are moving in the same direction. (d) It shows the normal situation that $R_4 < A$.

Response

Thank you for your careful comments. Indeed as you said, the fairness of the experimental comparison algorithm is very important. The experimental setup for all our comparison algorithms and our proposed algorithm is as follows:

1) We are conducting experiments on a unified experimental platform The overall experiments are conducted on a workstation with Intel Xeon(R) CPU E5-2650 v4 @ 2.20GHz NVIDIA TITAN Xp GPU. The proposed tracker and comparison algorithm are implemented on PyTorch deep learning platform.

2) In the comparison algorithm experiments, we compare multiple algorithms on multiple datasets, among which RSSRAI Data Set and UAV123 Data Set are public datasets.

Among these comparison algorithms, DCF, SAMF and DLSSVM are traditional algorithms with open source codes. The experimental results in the manuscript are obtained by our testing on their official source code.

For algorithms involving deep learning such as LCT, ECO and HDT, we adapt the open source code to the test data set without changing the network structure, and obtain the experimental results by fair testing.

For algorithms ATOM, DIMP and our algorithms, their frameworks are similar. So we use the consistent initialization of exact same settings: We uniformly use GOT-10K and IPIU^T as the training data set. Input patches size are 72×72 pixels from image regions corresponding to 5 times the estimated target size. In each epoch, 60% samples are randomly selected from GOT 10K and the remaining

40% from IPIU^T. We sample image pairs from each sequence with a maximum interval of 30 frames uniformly, and each batch includes 26 image pairs. The learning rate are also same as 2×10^{-4} in backbone network and 1×10^{-3} in IoU predictor learning with the same training epochs.

3) We use the same tracker evaluation indicators to conduct fair comparisons of multiple data sets on the same platform.

As mentioned above, we try our best to ensure fairness in comparative experiments to more objectively verify the performance of our algorithm.

Thank you very much for your valuable suggestion again! In fact, your suggestions give us a lot of inspiration.

Additional Questions:

Summary of Evaluation: Good

Organization: 4

Clarity: 4

Length: 3

References: 4

Correctness: 4

Significance: 4

Originality: 5

Attachments:4

If Survey Coverage:4

Contribution: 4

Please make very detailed technical and editorial comments and suggestions in your comments. If it is necessary to provide mathematical corrections, please email them to us as a pdf file. If you must get other information back to us that cannot be sent via email, please mail it to us. Your comments are an invaluable aid to the author to help in improving the overall technical quality, utility, and readability of the material. Such comments are not just useful, they are necessary to maintain the quality of the articles that are published in the SMC Transactions. Particular attention should be given to details that guide possible revisions, or that clearly explain reasons for rejection.:

What are the contributions of the paper?:

This manuscript proposes a collaborative learning tracking network for remote sensing video. It mainly includes three parts, the authors first propose a consistent receptive field parallel fusion module (CRFPF) to deal with the challenge of small goals, then propose a dual-branch spatial-channel co-attention (DSCA) module, to uses the spatial-channel co-attention mechanism to collaboratively learn the relevant information of the Target Classifier and IoU Regression, finally use geometric constraint retrack strategy (GCRT) to trace back the results of previous frames to re-evaluate and correct the tracking results. In my opinion, the manuscript is well written and clearly understandable. The most attractive feature of this manuscript is its novelty, and theobtained results are convincing.

What are the additional ways in which the paper could be improved?: I have some questions and suggestions as follows.

- For the challenge of small object, the current mainstream methods usually use feature pyramid for feature fusion among different feature levels. Why do the authors use image pyramid in input? What are the unique differences and advantages compared with the feature pyramid? In addition, what is the significance of keeping the consistent receptive field on the same level layer for all branches?
- Why do the authors use spatial co-attention in the Target Classifier branch, while use channel co-attention in the IoU Regression branch? Can these two different attention modules be exchanged between branches? If not, please give reasons.

- In Figure 4, why are the settings of attenuation functions designed to be different? What's the difference between the two? Which is more relaxed? Why is it so designed?
- Is the angle range in the geometric constraint limited to a small acute angle In Figure 5? Is there an obtuse angle?
- In the verification experiment, whether the initialization of several comparison algorithms is consistent or not, note that the fairness of the experimental comparison algorithm is very important, please make further elaboration.

2 Response to Reviewer 2

Comments:

The authors propose a collaborative learning tracking network for remote sensing videos, including a consistent receptive field parallel fusion module (CRFPF), dual-branch spatial-channel co-attention (DSCA) module, and geometric constraint re-track strategy (GCRT).

However, there are some issues to be addressed.

Reviewer 2 Comment 1

In multi-resolution sampling, why does image patch K choose 3? Because the value of K is very important for the extraction of depth features, is there a need for parameter evaluation here?

Response

Thank you very much for your opinion. Indeed, K is worth choosing in feature extraction. Theoretically, the larger the value of K, the richer the extracted depth features. But at the same time it also bring about a multiplied increase in parameters and affects the speed of convergence.

We make some roughly selection of parameters for K, but they are not systematically presented in original manuscript. So according to your comment, we think it is very necessary to do a detailed parameter estimation of value K. The specific experimental design is as follows. In the experiments, the detailed structure of each branch $\mathbb{B}_{I_k}$ is shown in Table 1 (please see the next page). Note that, when K takes a certain value, the overall network contains branches from $\mathbb{B}_{I_0}$ to $\mathbb{B}_{I_K}$. Considering that our datasets have many objects of size around 5×8 pixels, we conduct an overall test on these data, and the experimental results are shown in the following Table 2 (please see the next page):

It can be seen that, at the beginning, the accuracy of the network and computational complexity both increase with increasing K value. When $K = 3$, the accuracy of the network has risen to a stable level. We consider that the network at this time has been able to sufficiently extract the deep features of small objects. When $K > 3$, the accuracy of the network is no longer improved, but the computational complexity is still multiplying. Therefore, according to the object size relationship in the experiments and considering the balance between the parameter amount and accuracy, $K = 3$ is suitable for our application scenario tasks. We choose $K = 3$ as the number of branches in the paper.

In fact, the value of K does not have to be fixed at 3. Its value can be adjusted according to the specific practical application. We believe that its value may be affected by factors such as the size of the object itself, the magnitude of the training dataset, and timeliness requirements. For the size of the object, if the object is small, K does not need to be very large to satisfy the mining of deep features. An excessively large K may not necessarily bring about gains in accuracy and may also cause network redundancy; If the amount of training data is not large enough, it may not be able to support the training of many branches caused by too large K; For the timeliness of the task, the user needs to balance the value of K between the accuracy and timeliness according to the training time requirements of the task.

Your comments are very valuable, we have added the experiments and corresponding analysis of K to Section III C) of the manuscript.

Table 1: The specific structure of each branch $\mathbb{B}_{I_k}$ in CRFPF-Module.

Branch	Input Shape	Output Shape	Layer	r_*^*	*stride*	*padding*
$\mathbb{B}_{I_0}$	(36,36,3)	(18,18,64)	Conv[7×7,64]	1	2	1
	(18,18,64)	(18,18,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	1	1	1
$\mathbb{B}_{I_1}$	(72,72,3)	(36,36,64)	Conv[7×7,64]	2	2	3
	(36,36,64)	(18,18,64)	Pool[3×3]	–	2	1
	(18,18,64)	(18,18,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	1	1	1
$\mathbb{B}_{I_2}$	(144,144,3)	(72,72,64)	Conv[7×7,64]	3	2	3
	(72,72,64)	(36,36,64)	Pool[3×3]	–	2	1
	(36,36,64)	(36,36,128)	$\text{Conv}\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 2$	2	1	1
	(36,36,128)	(18,18,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	1	2	1
$\mathbb{B}_{I_3}$	(288,288,3)	(144,144,64)	Conv[7×7,64]	7	2	3
	(144,144,64)	(72,72,64)	Pool[3×3]	–	2	1
	(72,72,64)	(72,72,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	4	1	1
	(72,72,64)	(36,36,128)	$\text{Conv}\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 1$	2	2	1
	(36,36,128)	(18,18,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	1	2	1
$\mathbb{B}_{I_4}$	(576,576,3)	(288,288,64)	Conv[7×7,64]	19	2	3
	(288,288,3)	(144,144,64)	Pool[3×3]	–	2	1
	(144,144,64)	(144,144,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	9	1	1
	(144,144,64)	(72,72,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	4	2	1
	(72,72,64)	(36,36,128)	$\text{Conv}\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 1$	2	2	1
	(36,36,128)	(18,18,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	1	2	1
$\mathbb{B}_{I_5}$	(1152,1152,3)	(576,576,64)	Conv[7×7,64]	47	2	3
	(576,576,3)	(288,288,64)	Pool[3×3]	–	2	1
	(288,288,64)	(288,288,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	23	1	1
	(288,288,64)	(144,144,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	11	2	1
	(144,144,64)	(72,72,128)	$\text{Conv}\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 1$	5	2	1
	(72,72,128)	(36,36,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	2	2	1
	(36,36,128)	(18,18,64)	$\text{Conv}\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 1$	1	2	1

Table 2: Performance comparison with different value of K.

Values of K	Target Classier Loss	IoU Loss	Total Loss	Time
K=1	0.656	0.183	4.121	1h 27min
K=2	0.648	0.172	4.059	2h 13min
K=3	**0.639**	**0.172**	**3.997**	**3h 3min**
K=4	0.669	0.182	4.119	5h 51min
K=5	0.679	0.192	4.275	8h 32min

Reviewer 2 Comment 2

The value of i in formula (4) should be: i=2,...,L.

Response

Thank you very much for your careful comments. After careful inspection, we find that there are indeed some omissions in the expression of formula (4). We are very sorry for not giving an accurate representation of the value of i. We have corrected and rewritten the formula following your comments as:

In fact, the upper part of the formula represents the case where $i = 1$, and the lower part represents the case where $i >= 2$. So following your suggestion, we rewrite the formula into the following form:

$$\mathfrak{F}_k^i = \begin{cases} r_k^i \times (\mathcal{K} - 1) + 1 & , \quad i = 1 \\ \mathfrak{F}_k^{i-1} + [r_k^i \times (\mathcal{K} - 1) + 1] \cdot \prod_{p=1}^{i-1} stride_k^p & , \quad i \geq 2 \end{cases} \tag{4}$$

We have carefully checked other formulas in the manuscript to ensure that similar low-level mistakes do not recur. Thanks for your reminder to make our manuscript more rigorous.

Reviewer 2 Comment 3

What does the dot in formula (4) (5) represent? This should be multiplication, which needs to be explained.

Response

Thank you very much for your careful comments. We feel sorry for not make special explanation for the meaning of the dot in formula (5) (6) (i.e. formula (4)(5) in the manuscript). Indeed, as you said, the dot in the formula represent multiplication.

Specifically, in formula (5), the point represents the multiplication of two values, and in formula (6), the point represents the multiplication of the value and the matrix. They are all dot multiplications.

We follow your suggestion and add supplementary explanation after the formula in our manuscript as follow:

1) For the formula (5) (i.e. formulas (4) in the manuscript):

The receptive field $\mathfrak{F}_k^i$ of the i-th layer ($i = \{2, \cdots, L\}$) is shown in following recurrence formula as formula (5), where the dot represent dot multiplication operations in maths.

$$\mathfrak{F}_k^i = \begin{cases} r_k^i \times (\mathcal{K} - 1) + 1 & , \quad i = 1 \\ \mathfrak{F}_k^{i-1} + [r_k^i \times (\mathcal{K} - 1) + 1] \cdot \prod_{p=1}^{i-1} stride_k^p & , \quad i \geq 2 \end{cases} \tag{5}$$

2) For the formula (6) (i.e. formulas (5) in the manuscript):

Therefore, we propose an adaptive feature fusion way to obtain the final fused feature F_s as follows.

$$\begin{aligned} F_s &= \sum_{k=0}^{K} \lambda_k \cdot F_k^L \\ s.t. \quad \lambda_k &= \frac{\beta_k}{\sum \beta} \end{aligned} \tag{6}$$

where λ_k represents the fusion weight of feature F_k^L, which is normalized by β_k to interval $[0, 1]$. *And the dot represent dot multiplication operations.*

Thanks again for your careful comments!

Reviewer 2 Comment 4

There are many symbols in this paper, which leads to a hard follow for readers. For example, in Fig. 3, C should represent the number of bands, but in Fig. 2, C() represents convolution. I suggest that some parameters can be changed to more understandable statement, and re-phrase the whole paper to give a clearer description for each module.

Response

Thank you very much for your careful comments, and we are very sorry for our unclear notation. We have performed a comprehensive review of the entire manuscript, and have corrected the repetitions and misuse of symbols. For example, the specific changes as follows:

1) In Fig. 3 (Fig. 2 in manuscript), we correct the represent 'C' to '$Conv$', so that the repetitions of convolution are consistent in the manuscript.

2) In Fig. 3 (Fig. 2 in manuscript), we change the represent of the input RGB image patch $\{R_0^0, R_1^0, \cdots, R_k^0, \cdots, R_K^0\}$ to $\{I_0^0, I_1^0, \cdots, I_k^0, \cdots, I_K^0\}$. The symbols involved in the text have also been changed uniformly.

In this way, the symbolic representation of the image input can be kept in a unified correspondence with the input in below Fig. 4 of manuscript.

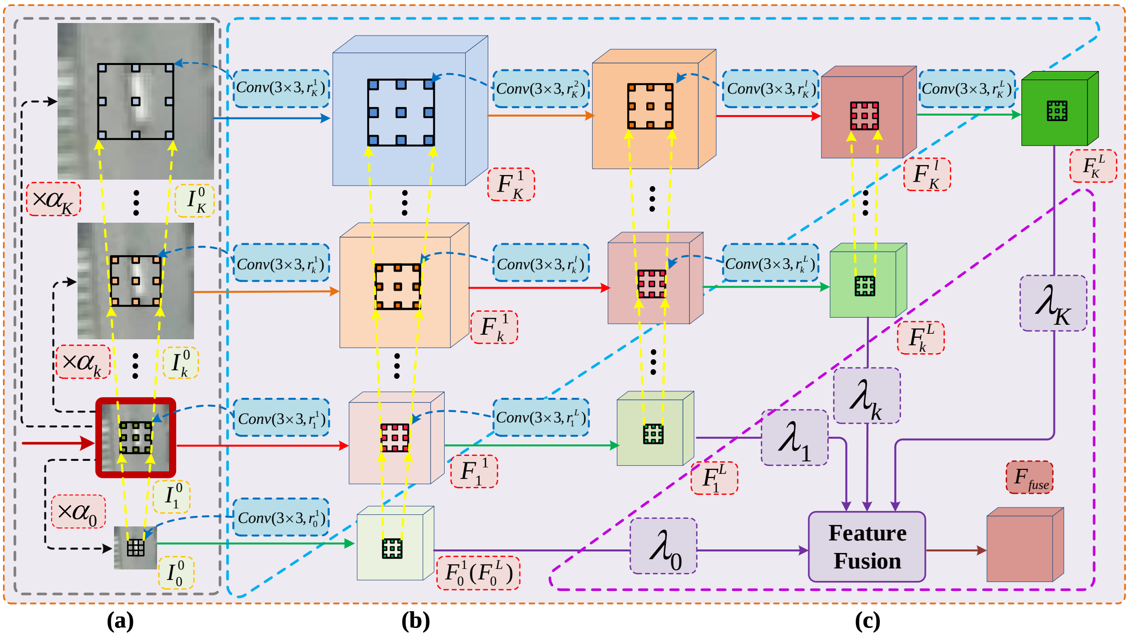


Figure 3: Flowchart of CRFPF-Module. The input is I_1^0 and the output is fused feature F_{fuse}. Construction of Parallel Multi-Resolution Input is in the grey dashed circle. Feature Extraction with Consistent Receptive Field is in the blue dashed circle. Adaptive Feature Fusion are in the purple dashed circle.

3) In Eq. (2)(3)(4) (in manuscript), we replace $\mathcal{K}$ (representing the convolution kernel size) with Ker to avoid confusion with the letter K representing the number of branches.

4) In Eq. (10) (in manuscript), we replace the letter A (representing a binary label) to letter G. The corresponding $a \in A$ is also replaced by $g \in G$. This avoids confusion with the letter A in Eq. (18) introduced in geometric constraint re-track strategy module.

5) For the simplicity and clarity of Eq. (11) in manuscript, we remove the subscript i of the letters ω_i, y_i μ_i. The revised equation becomes:

$$L_{cls}(X_S^C, Y) = \sum(f(x_s^c; \omega) - y)^2 + \mu\|\omega\|^2 \tag{7}$$

where, Y is the Gaussian sampling label centered at the object, X_S^C is the final attention feature, $y \in Y$ and $x_s^c \in X_S^C$. The value range of y is $[0, 1]$. ω is the weight of the fully convolutional layers, μ is the amount of regularization on ω.

6) Similar to the situation in 3), in order to avoid misunderstanding caused by the use of the letter K, we replace $X_C^0 \in \mathbb{R}^{K\times K\times C}$ with $X_C^0 \in \mathbb{R}^{J\times J\times C}$ (at line 320 and 326, page 5). That is to use the letter J instead of letter K to represent the spatial size of the feature.

7) The letter B of $\{B_{I_0}, \cdots, B_{I_k}, \cdots, B_{I_K}\}$ representing parallel convolution branches in consistent receptive field parallel fusion module is easy to be confused with the letter B representing region within bounding boxes in channel co-attention module and geometric constraint re-track strategy

module (E.g. B_g, B_e, B^t and B^t_{org}). Therefore, we use the different format $\mathbb{B}$ to represent parallel branches, that is, $\{\mathbb{B}_{I_0}, \cdots, \mathbb{B}_{I_k}, \cdots, \mathbb{B}_{I_K}\}$.

Following your comments, we try our best to correct unclear expressions in each module to give a clearer description. Corrections not mentioned above are marked in the manuscript one by one.

Reviewer 2 Comment 5

There are many grammatical errors in the paper, please check the full text in detail.

Response

Thank you very much for your careful comments. We apologize for a lot of low-level mistakes in English usage that reduce the readability of the paper and cause inconvenience. We try our best to correct the language expressions in the manuscript. In addition, we also consult several experts English-speaking experts in the field. Some modifications are shown as follows:

1) For sentence (Line 36, page 1): Object tracking is one of the key technologies in video analysis and understanding applications and understanding applications, and many advanced trackers for natural videos have been proposed.

we corrected it to: Object tracking is one of the key technologies in video analysis and understanding applications. Many advanced trackers for natural videos have been proposed.

2) For sentence (Line 122, page 2): For the tracked results, we design a re-track strategy to judge whether the objects is false tracked through geometric constraint,

we corrected it to: For the tracked results, we design a re-track strategy to judge whether the objects are false tracked through geometric constraint.

3) For sentence (Line 269, page 4): We design the spatial co-attention map S^r to measure the similarity of Z_S^1 and X_S^1 in corresponding spatial position.

we corrected it to: We design the spatial co-attention map S^r to measure the similarity of Z_S^1 and X_S^1 in the corresponding spatial position.

4) For Sentence (Line 357, page 5): The indiscernible appearance of objects leads to the difficulty to retrieve objects once lost.

we corrected it to: The indiscernible appearance of objects leads to the difficulty of retrieving objects once lost.

Following your suggestions, we have split and rewritten the long sentences to express more clearly. For confusing sentences, we have rearranged and expressed them in a way that is easier to understand. Corrections not mentioned above are also marked with red font in the revised manuscript. Thanks again for your valuable comments.

Reviewer 2 Comment 6

The network designed in this paper involves a lot of parameters. How to deal with the computational complexity in the training process?

Response

Thank you very much for your valuable comments.

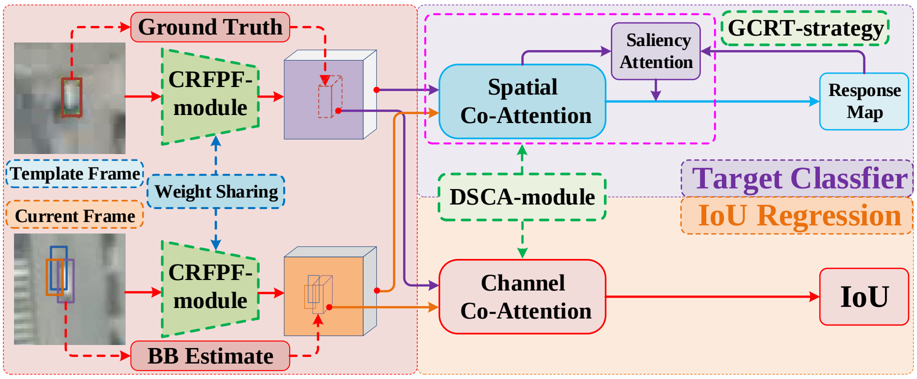


Figure 4: The overall process framework, including a consistent receptive field parallel fusion module (CRFPF), a dual-branch spatial-channel co-attention module (DSCA), and geometric constraint re-track strategy (GCRT).

Indeed, compared to the single-branch network framework, our multi-branch network has slightly more parameters. So we train the overall framework in a step-by-step manner to deal with the computational complexity. We follow the overall flow in Fig. 4 (Fig. 1 in manuscript) to explain our training process in detail:

1) Train the backbone network for feature extraction in CRFPF-module :

To construct the input to the network, we select two frames from the same video sequence in the training set within a fixed frame interval. For the pair of selected frames, we crop out the image patches that with several times (5 times in our experiments) the size of the object and centered on the bounding box. And we send them into the backbone network, that is CRFPF-module, with sharing weight. For the convenience of training, we add a classifier head after the backbone network.

According to quantitative calculation, the total number of parameters of the backbone network composed of parallel branches (when $K = 3$) is about 9.27 million, and the memory occupied is about $35.35MB$. The memory of other backbone network commonly used in remote sensing tracking like ResNet is $21.23MB$.

2) Train the Target Classfier and IoU Regression with Co-Attention in DSCA-module.

After the above backbone network training is completed, the parameters of this part are frozen in later training. Following, we train the subsequent Target Classfier and IoU Regression. Fused features from the above backbone network (that is the purple and the orange cubes) are delivered to dual co-attention module to obtain the classification response map and the IOU value of proposal regions. A trained Target Classfier can classify the feature map into target and background. Then the IoU Regression around the target to get an accurate tracking bounding boxes. At this point, the offline training process of the tracker is completed. And the number of our training sets is also sufficient to support the above training process.

According to our calculation, the total number of parameters of the Target Classfier branch with spatial co-attention module is about 2.52 million, and the memory is about $9.61MB$. The total number of IoU Regression branch parameters with channel co-attention module is about 2.14 million, and the memory is about $8.15MB$. Since there are not many layers and channels set for each network block, the amount of parameters is not very large as it looks.

3) In the testing process, the template frame (which is the first frame) and the current frame, are sent into the above trained network in pairs to obtain the preliminary tracking results. At this time, GCRT-strategy judges the result of the tracking box. The geometric constraints are used to measure whether the tracking result is abnormal. If the re-tracking procedure is triggered, the saliency template is updated to retrieve the mistakes and the object will be re-tracked.

We greatly appreciate your opinion and add the analysis of network complexity with Table VII in revised manuscript. Indeed, it is found during the experiment that the network design of our algorithm is relatively complex. In our future research, we are preparing to further simplify the network structure to improve the efficiency of the tracker while maintaining the accuracy of the algorithm.

We have added the training procedure of the algorithm and the calculation of the amount of parameters, as well as the corresponding analysis, to the revised manuscript.

Reviewer 2 Comment 7

Fig. 2 is not explained in the paper. In particular, the meaning of each step in this figure should be marked with respect to the statement.

Response

Thank you very much for your useful suggestions. In the description of Consistent Receptive Field Parallel Fusion Module, we only focus on the structure and principle, but neglect the detail steps of the overall process. This caused confusion in understanding. Therefore, we complement the detail introduction of this module completely with reference to Fig. 3 (Fig. 2 in manuscript). The specific steps are as follows:

1. The first step of the process is the area within the gray dashed circle.

(a) The original image patch I_1^0, enhanced by the red border, is the input of the module.

(b) Then a series of sampling rates $\{\alpha_0, \cdots, \alpha_k, \cdots, \alpha_K\}$ are calculated according to formula (1) (in manuscript).

(c) Finally, we perform multi-resolution sampling on the original image patch I_1^0, and obtain a series of image patches $\{I_0^0, \cdots, I_k^0, \cdots, I_K^0\}$ as inputs for multiple parallel branches.

2. The second part is shown in the blue dotted line in Fig. 3.

(a) First, we build a standard convolutional network branch $\mathbb{B}_{I_1}$ as the feature extraction branch of the initial image I_1^0.

(b) Then, according to the principle of receptive field consistency set by formula (2)-(4) (in manuscript), we obtain multiple parallel branches $\mathbb{B}_{I_k}, (k = \{0, 1, \cdots, K\})$ with given sampling rate $\alpha_k, (k = \{0, 1, \cdots, K\})$ and dilation rate r_1^1 of branch $\mathbb{B}_{I_1}$.

(c) Finally, the multiple resolution images patches $\{I_0^0, \cdots, I_k^0, \cdots, I_K^0\}$ are sent to their respective branches for feature extraction. So far, a series of hierarchical features $\{F_0^L, \cdots, F_k^L, \cdots, F_K^L\}$ are extracted.

3. The third step is the area within the purple dashed circle.

We adaptively fuse the hierarchical features $\{F_0^L, \cdots, F_k^L, \cdots, F_K^L\}$ obtained in 2) according to the fusion function expressed by formula (5)-(6) (in manuscript). The fusion feature F_{fuse} is obtained.

Indeed, as you said, the three subsections in our original manuscript are not intuitive enough to understand. Therefore, we modify the title of the subsection in the original manuscript as follows:

1) Construction of Parallel Image Pyramid Input.

2) Feature Extraction with Consistent Receptive Field.

3) Adaptive Feature Fusion.

At the same time, we describe Fig. 3 (Fig. 2 in manuscript) in detail, and add the corresponding explanation in the manuscript (e.g., Line 160-165, page 2; Line 206-215, page 3), so as to facilitate the reader's overall understanding of the module.

Thank you very much again. Your valuable comments are very meaningful and make our manuscript a greatly improvement.

Additional Questions:

Summary of Evaluation: Fair

Organization: 3

Clarity: 2

Length: 3

References: 4

Correctness: 3

Significance: 3

Originality: 3

Attachments:4

If Survey Coverage:4

Contribution: 3

Please make very detailed technical and editorial comments and suggestions in your comments. If it is necessary to provide mathematical corrections, please email them to us as a pdf file. If you must get other information back to us that cannot be sent via email, please mail it to us. Your comments are an invaluable aid to the author to help in improving the overall technical quality, utility, and readability of the material. Such comments are not just useful, they are necessary to maintain the quality of the articles that are published in the SMC Transactions. Particular attention should be given to details that guide possible revisions, or that clearly explain reasons for rejection.:

What are the contributions of the paper?:

The authors propose a collaborative learning tracking network for remote sensing videos, including a consistent receptive field parallel fusion module (CRFPF), dual-branch spatial-channel co-attention (DSCA) module, and geometric constraint re-track strategy (GCRT).

What are the additional ways in which the paper could be improved?:

- In multi-resolution sampling, why does image patch K choose 3? Because the value of K is very important for the extraction of depth features, is there a need for parameter evaluation here?
- The value of i in formula (4) should be: i=2,...,L.
- What does the dot in formula (4) (5) represent? This should be multiplication, which needs to be explained.
- There are many symbols in this paper, which leads to a hard follow for readers. For example, in Fig. 3, C should represent the number of bands, but in Fig. 2, C() represents convolution. I suggest that some parameters can be changed to more understandable statement, and re-phrase the whole paper to give a clearer description for each module.
- There are many grammatical errors in the paper, please check the full text in detail.
- The network designed in this paper involves a lot of parameters. How to deal with the computational complexity in the training process?

- Fig. 2 is not explained in the paper. In particular, the meaning of each step in this figure should be marked with respect to the statement.

3 Response to Reviewer 3

Comments:

The paper is well written and the three proposed modules are novelty and effective. Extensive experiments have been conducted to verify the effectiveness of proposed modules in solving three challenging problems for remote sensing videos, i.e detecting small-size objects, distinguishing the objects and background, and eliminating interference of similar objects.

Reviewer 3 Comment 1

But if the authors can discuss the time complexity analysis of their method, it will be perfect.

Response

Thank you very much for your valuable comments. As you said, I am very sorry that we only count the test time of each algorithm, but ignore the detailed calculation and analysis of its time complexity. As you said, computing and analyzing the time complexity of the proposed algorithm is an important measure to evaluate the algorithm. Here, taking convolution as an example, the calculation formula of time complexity is as follows:

$$Time \sim O(\sum_{l=1}^{D} M_l^2 \cdot K_l^2 \cdot C_{l-1} \cdot C_l) \tag{8}$$

where D represents the overall number of layers of the network; M_l is the side length of the output feature map of the l-th layer; K_l is the side length of the convolution kernel of the l-th layer; C_{l-1} is the number of input channels of the l-th layer, which is also that of the output channels of the $(l-1)$-th layer; C_l is the number of output channels of the l-th layer.

Combining the network structure in Table II in the manuscript, we can calculate the time complexity of the proposed algorithm and the compared baseline algorithm ATOM as shown in Table 3. It can be seen from the results that the complexity of the proposed algorithm is generally higher than that of ATOM.

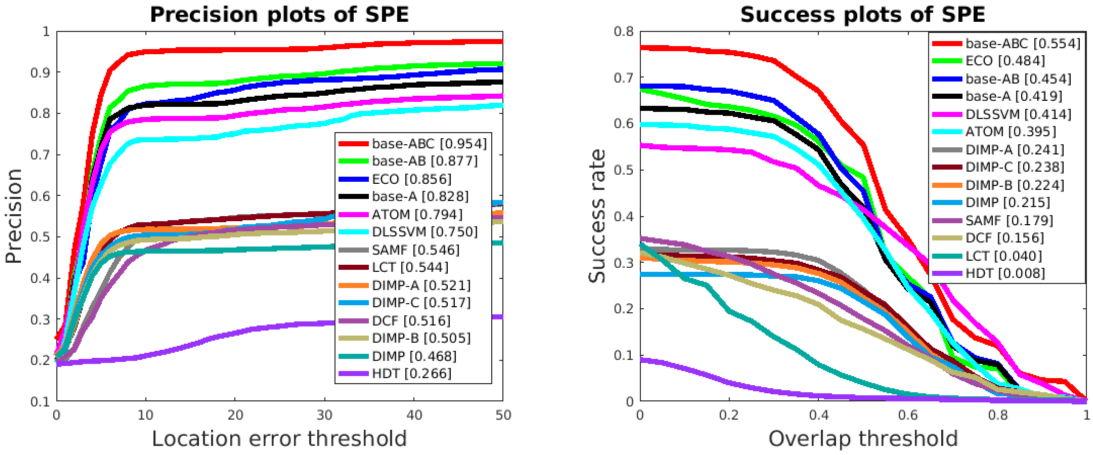
We analyze the reason: Compared with the single-branch backbone network of ATOM, the backbone network of the proposed algorithm is multi-branch, but the parameter quantity of each branch is less than that of ATOM branch. As shown in Table 3, the whole time complexity of ATOM is $O(13.67E+09)$, while the whole time complexity of the proposed algorithm is $O(18.91E + 09)$. Among them, the backbone network time complexity of ATOM is $O(9.95E + 09)$, and the backbone network time complexity of the proposed algorithm is $O(15.16E + 09)$, which is 1.52 times that of ATOM.

It can be seen that their parameters are in the same order of magnitude. In addition, the proposed algorithm also introduces an co-attention mechanism and a re-tracking mechanism, which also takes up a certain time complexity.

Thank you very much for your valuable comments, we add the calculation and analysis of time complexity in the corresponding position of the manuscript (line 725-733, page 11). In the near future, we will do further research on the algorithm, aiming to improve the tracking accuracy while simplifying the network, thereby reducing the time complexity of the framework.

Table 3: Model Complexity Analysis.

		Time Complexity		Space Complexity	
		Backbone	Total	Backbone	Total
ATOM	Parameters	$O(9.95E + 09)$	$O(13.67E + 09)$	$T(5.57E + 06)$	$T(11.16E + 06)$
	Memory	$37.05GB$	$50.94GB$	$21.23MB$	$42.58MB$
base-ABC (Ours)	Parameters	$O(15.16E + 09)$	$O(18.91E + 09)$	$T(9.27E + 06)$	$T(13.92E + 06)$
	Memory	$56.49GB$	$70.44GB$	$35.35MB$	$53.11MB$



(a) (b)

Figure 5: Comparison of different algorithms on IPIU data set. (a) Precision plot. (b) Success plot.

Reviewer 3 Comment 2

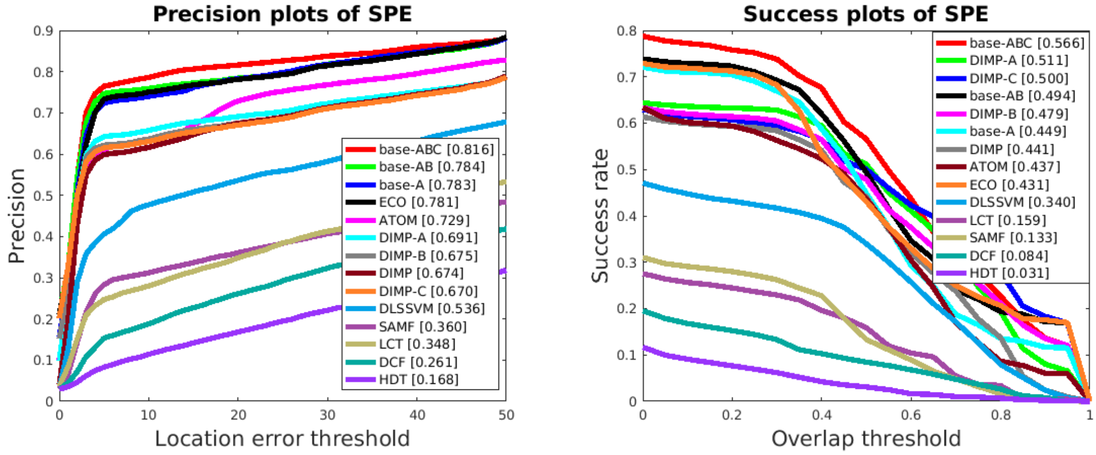
Besides, I am curious about whether the proposed three modules can also improve the performance of other comparative methods, not only ATOM. It will be more convinced to verify whether the three modules are model-agnostic.

Response

Thank you very much for your constructive comments. Indeed as you said, the original experiments in our manuscript only evaluated the performance of the proposed three modules with ATOM as the baseline. We are also curious about the performance of our proposed module on other comparative algorithms. Therefore, we select the deep learning tracker DIMP as an additional baseline, and respectively embed CRFPF-Module, DSCA-Module and GCRT-Strategy into it for supplementary experiments, to further verify the effectiveness and universality of each module.

For brevity in the experiment, CRFPF-Module embedded in DIMP is denoted as DIMP-A, DSCA-Module embedded in DIMP is denoted as DIMP-B, and GCRT-Strategy embedded in DIMP is denoted as DIMP-C. Therefore, the overall experimental results after supplementation are as follows:

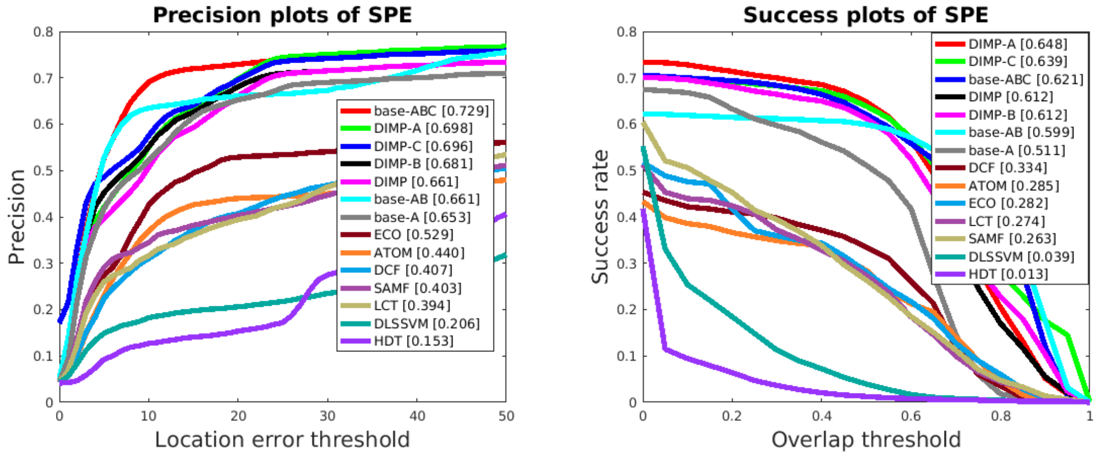
As shown in Table 4 (Table VI in the manuscript) and Fig. 5-7 (Fig. 12 in the manuscript), overall in the three datasets, DIMP-A, DIMP-B, and DIMP-C have improvements in Success and Precision compared to the DIMP algorithm alone.



(a)

(b)

Figure 6: Comparison of different algorithms on RSSRAI data set. (a) Precision plot. (b) Success plot.



(a)

(b)

Figure 7: Comparison of different algorithms on UAV123* data set. (a) Precision plot. (b) Success plot.

Table 4: Ablation Study and Performance Comparison Results of Three Data Sets.

	IPIU			RSSRAI			UAV123*		
	Precision	Success	FPS	Precision	Success	FPS	Precision	Success	FPS
DCF [15]	0.516	0.156	428.782	0.261	0.084	338.482	0.407	0.334	455.448
SAMF [56]	0.546	0.179	20.995	0.360	0.133	16.963	0.403	0.263	24.813
DLSSVM [57]	0.750	0.414	100.319	0.536	0.340	83.592	0.206	0.039	104.636
LCT [58]	0.544	0.040	35.795	0.348	0.159	33.697	0.394	0.274	43.012
ECO [22]	0.856	0.484	15.167	0.781	0.431	13.136	0.529	0.282	16.760
HDT [23]	0.266	0.008	46.817	0.168	0.031	31.672	0.153	0.013	51.864
DIMP [27]	0.468	0.215	18.361	0.674	0.441	14.733	0.661	0.612	22.612
DIMP-A	0.521	0.241	22.541	0.691	0.511	17.495	0.698	**0.648**	25.280
DIMP-B	0.505	0.224	17.230	0.675	0.479	14.933	0.681	0.612	20.957
DIMP-C	0.517	0.238	15.282	0.670	0.500	13.476	0.696	0.639	21.072
ATOM [26]	0.794	0.395	20.731	0.729	0.437	16.954	0.440	0.285	23.696
base-A	0.828	0.419	25.060	0.783	0.449	19.827	0.653	0.511	27.293
base-AB	0.877	0.454	16.103	0.784	0.494	14.932	0.661	0.599	19.692
base-ABC (Ours)	**0.954**	**0.554**	14.051	**0.816**	**0.566**	12.516	**0.729**	0.621	15.671

Specifically, DIMP-A is embedded with the parallel multi-resolution feature extraction branches, thereby obtains robust hierarchical fusion features for small objects, so the Precision and Success are significantly improved; for DIMP-B, under the guidance of the dual-branch co-attention mechanism, Target Classfier and IoU Regression are more accurate, so the tracking accuracy is improved more; for DIMP-C, tracking mistakes are reduces due to the introduction of the geometric constraint re-track strategy, therefore Success is improved.

Note that, on RSSRAI data set, Success of DIMP-C (i.e. 0.500) is improved compared with that of DIMP (i.e. 0.441), but its Precision (i.e. 0.670) slightly decreases by 0.004 compared with that of DIMP (i.e. 0.674). Since the resolution of the RSSRAI data set is 1.13m/pixel, the object blur is more serious than the other two data sets. In this case, when re-tracking strategy is used alone, the accuracy of bounding boxes obtained during regression is affected. In this way, although the target object can be recovered (Success is improved), precision of bounding boxes obtained from regression is affected.

In addition, on UAV123* data set, Success of DIMP-A is improved to 0.648 compared to baseline DIMP (0.612). Since this data set is shot on drone platforms, there are a lot of flips, target deformation, and angle changes of view. This is different from satellite remote sensing videos with small changes in the angle of view. The DIMP baseline algorithm has an online updated classifier module, so its algorithm performance on UAV123* data set is overall better than the baseline ATOM algorithm.

In summary, the three proposed modules can also be effectively transferred to other tracking algorithms and bring about performance improvements. This further demonstrates the extensibility of the module. Based on your comments, we add the above experimental content and experimental analysis to the experimental section in the manuscript.

Thanks again for your valuable suggestion to make our experiment more complete and convincing.

Reviewer 3 Comment 3

Although the proposed three modules have shown great advantages, DSCA-module needs the ground truth of the template frame by frame, which limits the flexibility of the strategy in practical application. In practice, it is hard and laborious to annotate the ground truth of each frame.

Response

Thank you very much for acknowledging our manuscript. Indeed, as you said, the training process of the DSCA-module relies on a large number of annotated tracking frames, which limits the flexibility of the strategy in practical application. Although with the emergence of more and more large-scale annotated public tracking datasets, this contradiction has been alleviated to a certain extent. But this is still one of the challenges faced by current deep learning trackers.

During testing, in order to use as little data as possible to achieve effective tracking, in fact, we only use the ground truth of the first frame as template to guide the learning of the attention mechanism.

In the online tracking process, the tracking result of the current frame is obtained from the pseudo-labels of the previous frames and the ground truth of the template frame. And the tracking result of this frame is used as a pseudo-label to predict the next frame.

Thank you for this very instructive comment. In the future research, we will take this challenge as a meaningful direction to continue to improve the training and learning method of the framework, so that it can reduce the dependence on the amount of training data while maintaining the accuracy. Finally, we have supplemented the analysis and outlook in this direction in the conclusion section of the manuscript based on your comments.

Reviewer 3 Comment 4

Besides, the additional three modules plugin current model, e.g. ATOM, it will increase computational complexity and slow the speed of inference. In some practical scenarios, it will be limited by some low-computational devices, such as Unmanned Aerial Vehicle (UAV). All in all, the efficiency of their algorithm should be improved in the future.

Response

Thank you very much for your considerate comments. Indeed, computational complexity and algorithm efficiency are valuable directions for extending algorithm performance. The computational efficiency of our algorithm is indeed somewhat slower than the original model, even though they are in the same order of magnitude. When designing the model, we mainly focus on improving the tracking accuracy of small objects in remote sensing scenes, so we adopt a multi-branch parallel network structure, which increased the computational complexity of the network. In addition, the mechanism of re-tracking is added, which also limits the overall speed of the algorithm.

In the near future, we will further simplify the network structure to improve the efficiency of the tracker under the condition of ensuring the accuracy of the algorithm. For example, we are thinking of borrowing from the teacher-student network in distillation learning to lighten the structure of the test network, so that our tracking framework is also scalable on low-computational devices.

Correspondingly, we have added an analysis and outlook in this direction to the conclusion section of the manuscript.

Additional Questions:

Summary of Evaluation: Good

Organization: 5

Clarity: 5

Length: 3

References: 5

Correctness: 5

Significance: 5

Originality: 5

Attachments:1

If Survey Coverage:4

Contribution: 4

Please make very detailed technical and editorial comments and suggestions in your comments. If it is necessary to provide mathematical corrections, please email them to us as a pdf file. If you must get other information back to us that cannot be sent via email, please mail it to us. Your comments are an invaluable aid to the author to help in improving the overall technical quality, utility, and readability of the material. Such comments are not just useful, they are necessary to maintain the quality of the articles that are published in the SMC Transactions. Particular attention should be given to details that guide possible revisions, or that clearly explain reasons for rejection.:

What are the contributions of the paper?:

In this paper, the authors propose a collaborative learning tracking network for remote sensing videos, including a consistent receptive field parallel fusion module (CRFPF), dual-branch spatial-channel co-attention (DSCA) module, and geometric constraint re-track strategy (GCRT), to solve three challenging problems, i.e detecting small-size objects, distinguishing the objects and background, and eliminating the interference of similar objects respectively. The extensive experimental results on multiple data sets demonstrate the effectiveness of their proposed three modules.

What are the additional ways in which the paper could be improved?:

Although the proposed three modules have shown great advantages, DSCA-module needs the ground truth of the template frame by frame, which limits the flexibility of the strategy in practical application. In practice, it is hard and laborious to annotate the ground truth of each frame. Besides, the additional three modules plugin current model, e.g. ATOM, it will increase computational complexity and slow the speed of inference. In some practical scenarios, it will be limited by some low-computational devices, such as Unmanned Aerial Vehicle (UAV). All in all, the efficiency of their algorithm should be improved in the future.

4 The End

We tried our best to improve the manuscript and made some changes in the manuscript. These changes will not influence the content and framework of the paper. And here we did not list the changes but marked in red in revised paper. We appreciate for Editors/Reviewers' warm work earnestly, and hope that the correction will meet with approval.

Once again, thank you very much for your comments and suggestions.