A Real-time Anti-distractor Infrared UAV Tracker with Channel Feature Refinement Module*

Houzhang Fang†[1], Xiaolin Wang[1], Zikai Liao[1], Yi Chang[2], and Luxin Yan[3]

[1]Xidian University, Xi'an, China
houzhangfang@xidian.edu.cn, wxl@stu.xidian.edu.cn, lzk773629528@163.com
[2]Pengcheng Laboratory, Shenzhen, China
owuchangyuo@gmail.com
[3]Huazhong University of Science and Technology, Wuhan, China
yanluxin@hust.edu.cn

Abstract

The unmanned aerial vehicles (UAVs) have been widely used in various application fields, yet unauthorized use of UAVs raises great threats for restricted areas and public security. Therefore, it is urgently necessary to develop a practical anti-UAV target tracking technique. In this paper, we propose a real-time anti-distractor infrared UAV tracker for infrared anti-UAV tasks, which employs a global real-time perception mechanism to find candidate targets, then utilizes spatial-temporal information to obtain the real UAV target. Moreover, we integrate a channel feature refinement module into multi-scale feature fusion to better enhance the representation of the finer features of the UAV targets channel-wisely, thus improving the tracking performance. We test the performance of the proposed method and the other competitive ones on the constructed UAV dataset from ourselves, and eventually verify the validity of the proposed method as the best performing method with a better balance between tracking accuracy and speed.

1. Introduction

In recent years, unmanned aerial vehicles (UAVs) have been widely used in both civil and military areas. With the rapid development of low-altitude, slowly-moving and small-scale UAVs [1], uncooperative UAV is posing many potential threats towards public safety as well as aerial security [2, 3], urging the research on anti-UAV target tracking technologies [4]. Meanwhile, the infrared imaging technology, with its all-day imaging capability and various-weather working flexibility, has been one of the most important technologies for constant surveillance of UAV targets at long-range, and has already become a crucial complement to other UAV tracking technologies like radar [5]. However, there are several difficulties in the task of infrared anti-UAV target tracking: (1) In long-distance imaging, the UAV targets are relatively small in scale and weak in visual features, and thus it is difficult to extract discriminatory features [6–8]; (2) Infrared UAV tracking can be easily affected by complex backgrounds (*e.g.*, trees, buildings, heavy cloud, and strong clutter) and distractors (*e.g.*, birds); (3) Tracking failures may occur as UAV target shifts its locations between the two neighboring frames drastically due to its sudden motion or the instability of the infrared imaging platform; (4) In order to promote the practical application, anti-UAV tasks usually require the tracking algorithms to possess a real-time processing ability.

Up to now, many target tracking methods have been developed [9–29]. Target tracking algorithms locate the target largely by taking advantage of the spatial information of the targets as well as the temporal correlation within the sequences [9–11]. Traditional target tracking algorithms, such as KCF [12] and TLD [13], can poorly model the appearance of the target in complex scenarios, thereby yielding incorrect tracking results. As deep learning (DL) advances, a considerable number of DL-based visual tracking algorithms emerged, *e.g.*, MDNet [14], SiamFC [15], CFNet [16], and those DL-based algorithms can be generally divided into two categories: in the first category, the deep

*This work was supported in part by the National Natural Science Foundation of China under Grant 41501371 and Grant 61971460, and in part by the Open Research Fund of the National Key Laboratory of Science and Technology on Multispectral Information Processing under Grant 6142113190103.

†Corresponding author

learning network is utilized as a feature extractor embedded into the traditional target tracking algorithm [17–19]; in the second one, the network is trained end-to-end to directly output the tracking results [14, 20, 21]. Recently, target tracking algorithms based on the siamese network achieved breakthroughs in improving the balance between the tracking precision and the tracking efficiency [15, 22–28], where the network trains the similarity metric function offline with image pairs. SiamFC [15] converted the tracking task into a template matching task, which is simple and fast, but cannot properly deal with scale or distance variances of the UAV targets. SiamRPN [23] addressed this issue by introducing the Region Proposal Network (RPN) to its architecture, but still suffers from comparatively low tracking precision due to its limited modeling ability of the backbone network AlexNet [30]. Based on SiamRPN, SiamRPN++ [29] improved the tracking precision by adjusting the sampling method of the positive samples in the network training stage.

The aforementioned algorithms, classified as short-term target tracking methods, whose performance are highly subject to their insufficient ability to discriminate appearance model as well as search region limitation, and thus it is difficult for them to recapture the target after tracking failure, whereas long-term target tracking methods have the advantage for anti-UAV tasks in these aspects. Off-the-shelf long-term tracking algorithms can be coarsely divided into two classes: one, exemplified by TLD, has created the classic paradigm of combining local tracking with global detection; the other one is to search globally on the feeding image while having the local tracking method as an auxiliary temporal constraint. The state-of-the-art long-term DL-based target tracking algorithms follow these two tracking patterns. For instance, SPLT [31] opts to use SiamRPN as its basic tracker, and changes its searching strategy through the skimming and the perusal modules. GlobalTrack [32] and Siam R-CNN [33] leverage the successes from target detection, and incorporate them into Siamese network architecture, making great progress in the field of long-term target tracking. Nevertheless, when dealing with infrared UAV targets, some issues remain to be solved: (1) Current long-term target tracking algorithms struggle to process infrared UAV target images in a real-time fashion, thus being inapplicable in practical scenarios. (2) UAV targets in long-distance imaging are relatively tiny in scale, with barely any obvious features, making it extremely challenging to precisely distinguish their features, especially when the background contains trees, heavy cloud or strong clutter. (3) Birds and other distractors may lead to unstable tracking performance. Therefore, a tracking algorithm well-balanced between the tracking precision and the tracking efficiency is urgently needed for infrared anti-UAV target tracking tasks.

To solve the three problems above, inspired by GlobalTrack and Siam R-CNN, we refer to the regression-based single-stage target detector YOLOv3 [34]. We combine it with siamese architecture, and come up with a real-time anti-distractor infrared UAV tracking model named SiamYOLO, which obtains a better equilibrium between the performance and the efficiency of tracking. Besides, since channel attention mechanism can highlight the representative features channel-wisely [35], we integrate it into the multi-scale feature fusion operation, which enhances the model's ability to refine finer features of UAV targets while repressing non-UAV ones. Furthermore, from the spatial-temporal perspective, we locate the real UAV target from its tracklet by using the spatial-temporal information of candidate targets. This can discard distractors and thus improving the model's anti-distractor ability.

Our contributions are summarized as follows:

(1) We incorporate an efficient searching strategy based on single stage detector into the siamese network architecture, and construct a real-time global-search infrared UAV tracking model.

(2) We integrate the channel feature refinement module (CFRM) into multi-scale feature fusion, which enriches the representation of the finer features of the UAV targets channel-wisely.

(3) We introduce a spatial-temporal information-based anti-distractor module, which further judges the similarity between the distractors and the real target, and find it effective in improving model's anti-distractor ability.

2. Proposed method

In this paper, we study the feasibility and the performance of a basic idea towards anti-UAV task, which is to globally search for candidate targets in the feeding frame and to use spatial-temporal information to discard distractors. By following this tracking paradigm, especially in long-term tracking scenarios, we can attain a better tracking performance by devising a more accurate candidate target generator and a more robust motion model.

Specifically, we propose a real-time anti-distractor infrared UAV tracking model (*i.e*. SiamYOLO), and the overview of SiamYOLO is shown in Fig. 1. It mainly consists of a real-time global tracking component and a spatial-temporal information-based anti-distractor module, which is featured by the following four points.

(1) Classic long-term tracking algorithms adopt local tracking combined with the redetection after target disappearance, but may easily miss the UAV target due to distractions and occlusions from the background as well as its unpredictable flying route. Therefore, we use a global instance search-based tracker to search the target on a global scale, which overcomes the disadvantage from traditional long-term tracking patterns.

(2) To achieve real-time tracking while the search region

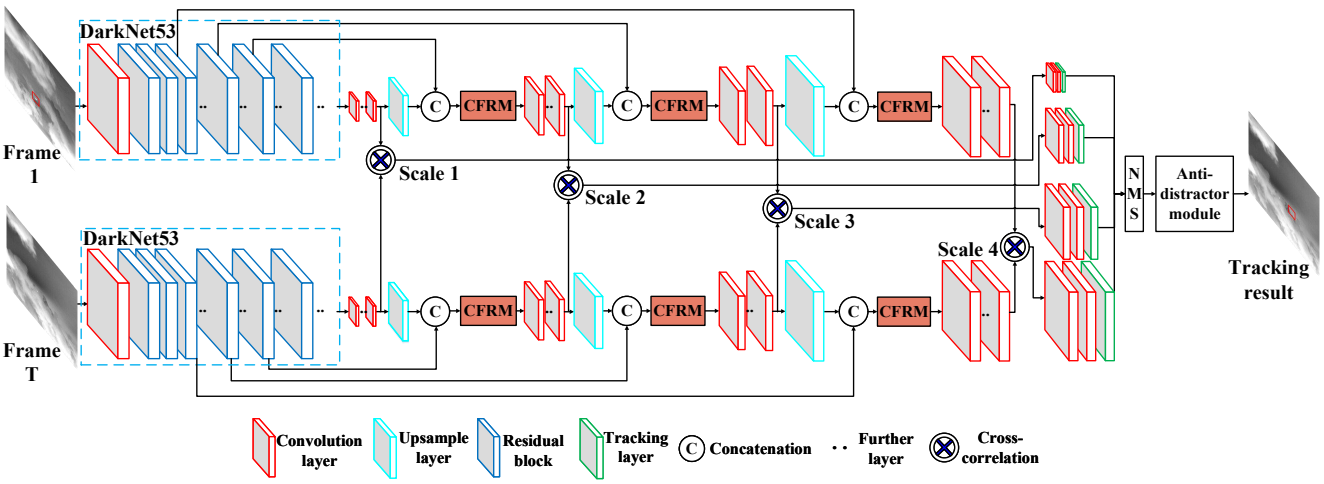

Figure 1: The overview of the proposed SiamYOLO. Our tracker combines YOLOv3 with the siamese network structure, which can effectively improve the real-time performance, and the search area is the whole image to solve the situation where tracking fails because of rapid movement of the UAV. The CFRM can enhance the features of UAV targets in complex scenarios and improve the tracking precision. In addition, anti-distractor module utilizes the UAV trajectory information from the space-temporal dimension, which improves the model's ability to discriminate UAV targets from similar distractors.

enlarges, we construct a real-time multi-scale infrared UAV tracker based on siamese network by referring to YOLOv3, a single-stage detector well-known by its speed superiority.

(3) For better feature extraction of UAV target from background and clutter, we embed CFRM in multi-scale feature fusion, where it can enhance the UAV target features while repressing the others, leading to the tracking localization and the robustness improvements.

(4) We use the spatial-temporal clues of the UAV target with the Kalman filter [36] to predict the target location based on current tracklets, compare the candidate targets with that predicted location and eventually retain the real UAV target as well as discarding the other ones. In contrast to merely calculating visual feature similarity, this approach is relatively more robust.

2.1. Real-time multi-scale infrared UAV tracking component

To accomplish global real-time tracking, we refer to YOLOv3 to establish our basic tracking component. Although we are not the first one to incorporate target detection mechanism into the target tracking task, our tracker still has the edge compared to other competitive methods. For instance, SiamFC and SiamRPN track targets located around the center of the image, which enforces the network to learn the location deviation from the image center, thereby yielding poor tracking results. However, our tracker is designed to track the target regardless of where the target is on the image, breaking the limits of spatial invariance of the network architecture. Additionally, our tracker can retrack and rectify the tracking results on the spatial-temporal dimension even if the tracking component tracks the distractors, showing its greater robustness than SiamRPN [23] and GlobalTrack [32].

The feature extraction backbone network of our tracking component is DarkNet-53 [34], from which we use the output feature maps of layer No. 12, 28, 44 and 53, and group them as a pyramid structure representing the multi-scale features of the input image. On one hand, the shallow feature maps are coarser in semantics but finer in resolution, which is conducive to small infrared UAV target localization; on the other hand, deep feature maps have bigger receptive fields and abundant semantic information, which benefit the UAV target classification.

The backbone network ϕ accepts the template image Z and the search image X as its two inputs, extracts their features and obtains the template features for the template image $f_{z\_i} = \phi(Z), i \in \{1, 2, 3, 4\}$, and the search region features $f_{x\_i} = \phi(X), i \in \{1, 2, 3, 4\}$, both of four multi-scale feature branches, respectively. After that, we clip out and calibrate those UAV target features f_{obj_i} of four different scales using region of interest (RoI) align [37] from template features using

$$f_{obj\_i} = \mathcal{R}(b_{obj}, f_{z\_i}), \tag{1}$$

where $\mathcal{R}$ represents the RoI align operation, b_{obj} the ground-truth bounding box, and $i \in \{1, 2, 3, 4\}$, $f_{obj\_1} \in \mathbb{R}^{512\times3\times3}$, $f_{obj\_2} \in \mathbb{R}^{256\times6\times6}$, $f_{obj\_3} \in \mathbb{R}^{128\times12\times12}$, $f_{obj\_4} \in \mathbb{R}^{64\times24\times24}$. Subsequently, we achieve the similarity between f_{obj_i} and f_{x_i} of the same scale by calculating

$$\hat{x}_i = \phi_z(f_{obj\_i}) \otimes f_{x\_i}, \tag{2}$$

where $\otimes$ indicates convolution, ϕ_z converting f_{obj_i} to an 1×1 convolutional kernel over f_{x_i}. By that, $\hat{x}_i$ and f_{x_i} have the same size, which facilitates the follow-up classification and bounding box regression of the tracking layer. The candidate target locations are obtained from the tracking layer and non-maximum suppression (NMS) based on $\{\hat{x}_i\}_{i=1}^{4}$.

2.2. Channel feature refinement module (CFRM)

The backbone network is able to extract features from the feeding image and gradually stack them along the channel dimension, so it is significantly crucial to fully utilize them for target discrimination. To this end, we propose to use the feature pyramid structure, fusing classification information from the deep layer and the localization ones from the shallow layer. However, we find it is relatively inaccurate to use the fused channel-wise features directly from the pyramid structure as the UAV target becomes small and the background is complicated, leading to mis-tracking or tracking failure. We conjecture that the feature pyramid structure has submerged the original channel-wise UAV target features with other redundant ones, and thus later operations struggle to determine the most contributing features.

To address this problem, we introduce the channel feature refinement module (CFRM) to our model. The overview of this module is shown in Fig. 2, which we embed at the fused feature maps from the feature pyramid structure. Our intention is to weigh the features channel-wisely by the contribution each channel-wise feature has to the infrared UAV target tracking. From Fig. 2, the module firstly takes an input image of size (C, H, W) and obtains the initial feature significance statistics of each channel sized $(C, 1, 1)$ using global average pooling (GAP). Then the statistics are forwarded to a convolutional sub-network to calculate $(C, 1, 1)$ relational feature significance statistics that are not channel-wisely mutually-exclusive. Eventually, those statistics are normalized by an upcoming sigmoid function, and are multiplied onto the input feature map channel-wisely.

By incorporating the CFRM, the model is characterized by an automatic channel feature weighting mechanism. The CFRM serves as a channel feature weighting branch that calculates the feature priority in each channel of the feature map, hence highlighting the infrared UAV target and suppressing other unimportant distracting features, and further leading to effectively discarding distractors and retaining the real infrared UAV target, especially in complex background and when the target is small, occluded or distracted. From this perspective, the model is capable of adaptively refining channel-wise features from feature submergence, thereby yielding more precise and robust results in the course of infrared UAV tracking, and ultimately improving the overall infrared UAV target tracking performance.

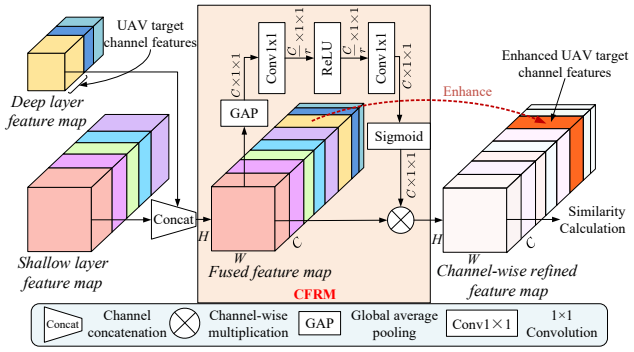

Figure 2: The illustration of the proposed channel feature refinement module (CFRM). The CFRM weighs the features channel-wisely by the contribution each channel-wise feature has to the infrared UAV target tracking.

2.3. Anti-distractor module based on spatial-temporal information

After capturing template-similar candidate targets using tracking component, it is rather computation-costly and time-consuming to discriminate two similar objects only by their visual features. Therefore, we propose a spatial-temporal information-based anti-distractor module over the spatial and the temporal dimensions to suppress the interference from distractors and background clutter. The main idea of this module is to exclude outliers in respect of continuous target movement, and the details are specified as follows (*c.f.* Algorithm 1). We firstly initialize the Kalman filter using the location of the template frame $templateBbox$. Subsequently, the tracking component generates candidate object locations $candidateBbox_t$ at t frame, and the Kalman filter predicts a target location based on the tracklets (line 1). After that, we calculate the distance between each candidate target and the predicted location, and select the one candidate target that has the minimum distance as the tracking result bounding box $Bbox$ (line 2). Then we compare this distance with the threshold parameter τ: if the distance is no greater than τ, the module will take the corresponding bounding box as the final tracking result and correct the Kalman filter by the bounding box Bbox (line 3-5), otherwise the module believes no UAV target exists in the image and returns a no-result descriptor (line 6-7). Empirically, we set τ to 50.

3. Experiments

In this section, we conduct some experiments using the proposed method along with some state-of-the-art long-term tracking methods based on our anti-UAV infrared dataset, and evaluate the performance of these methods. We firstly introduce the composition of our dataset, then present some implementation details of our experiments, and even-

Algorithm 1 Sift candidate targets at frame t

Input:

$candidateBbox_t$, τ

$\triangleright\tau$ is set to 50

Output:

$Bbox$ or Non_exist

$\triangleright Non\_exist$ is a descriptor of no result bounding box

$predictBbox \Leftarrow Kalman.predict()$;

$Bbox = \arg\min ||canBbox_i - predictBbox||_2$, where $canBbox_i \in candidateBboxes_t$;

if $||Bbox - predictBbox||_2 \leq \tau$ **then**

  $Kalman.correct(Bbox)$;

  return $Bbox$

else

  return Non_exist

end if

tually demonstrate the experiment results as well as the experimental analysis.

3.1. Dataset

We construct our Anti-UAV infrared dataset by collecting infrared images of UAV with our infrared imaging devices. In our perspective, the common difficulty of tracking infrared UAV targets is that, apart from their regular presence in dynamically-motioned complex backgrounds (*e.g.*, backgrounds with buildings and sky with heavy cloud, in which the infrared UAV target could be easily submerged), they often vary in shape and scale in the field of sight, hard to be spotted especially when being small, and easy to be mis-tracked due to distractors similar to the provided template.

In this work, we gathered around 14,700 images and use them as the training data. We then select another five sequences as our testing data to verify the chosen methods, and these two data are intersection-free. The details of the testing sequences are listed in Table 1, where Seq. 1 are the long-distance UAV target infrared images, Seq. 2 includes complex backgrounds such as strong cloud, electrical wires and buildings, Seq. 3 has a distractor, Seq. 4 contains trees, wires and cloud as the UAV target is extremely small in sight, and Seq. 5 is mainly about complex tree background.

3.2. Implementation details

We first use the training dataset to train the proposed network and tune a group of training hyperparameters for obtaining the best performance. The total training epochs, initial learning rate, momentum, weight decay, and batch size are set to 500, 0.001, 0.9, 0.0005 and 36, respectively.

Since the proposed method belongs to the long-term tracking category, we select TLD, SiamRPN-LT, SPLT, and GlobalTrack as the comparison methods. We use the same training dataset to train the comparison methods (except for TLD), and the training hyperparameters are tuned meticulously to the optimal. We conduct all the experiments on a server with 2.40 GHz Intel Xeon Silver 4210R CPU and three NVIDIA RTX3090 GPU cards. For software configurations, we use PyTorch 1.8.1 with CUDA 11.1.

3.3. Experiment results

3.3.1 Evaluation metrics

To fairly compare different tracking methods, we use Precision (P), Recall (R), F1 score ($F1$) [38] and tracking average accuracy (acc) [39] as our evaluation metrics. The former three metrics are commonly used in evaluating long-term tracking algorithm, and are related to the deviation of the tracking bounding box and the corresponding ground-truth box. The last metric is associated to the IoU (Intersection over Union) of those two boxes. The definition of P, R, $F1$ and acc are given as follows.

$$P = \frac{1}{N_p}\sum_{f=1}^{N_p}\sum_{t=1}^{M}\sqrt{\left(T_x^{f,t} - G_x^{f,t}\right)^2 + \left(T_y^{f,t} - G_y^{f,t}\right)^2}, \quad (3)$$

$$R = \frac{1}{N_g}\sum_{f=1}^{N_p}\sum_{t=1}^{M}\sqrt{\left(T_x^{f,t} - G_x^{f,t}\right)^2 + \left(T_y^{f,t} - G_y^{f,t}\right)^2}, \quad (4)$$

$$F1 = \frac{2 \times P \times R}{P + R}, \quad (5)$$

$$acc = \frac{1}{T}\sum_{t=1}^{T} IoU_t \times \mathbf{1}[v_t > 0] + p_t \times (1 - \mathbf{1}[v_t > 0]). \quad (6)$$

In the above formulas, N_p denotes the number of frames which the tracking method yields tracking results on, and N_g the number of frames which have any ground-truth labels. P and R are both averaged by N_p and N_g. M is the number of tracked result in one frame, usually set as 1 in single-target tracking tasks. Subscripts x and y in Eq. 3 and Eq. 4 indicate the coordinates of tracked target T and ground truth G, respectively.

In Eq. 6, IoU_t defines the Intersection over Union (IoU) between the tracked bounding box and its corresponding ground-truth bounding box. P_t equals 1 when the tracked bounding box is empty and 0 otherwise, and v_t is the ground-truth existence flag of the target. The Iverson bracket indicator function $\mathbf{1}[v_t > 0]$ equals 1 when $[vt > 0]$ and 0 otherwise. Finally, acc is averaged over T frames.

3.3.2 Comparisons

The quantitative comparison results are shown in Table 2. For better demonstration on method generalizability, we calculate the average value of all 5 test sequences on each evaluation metric. Red and blue represent the highest and the second highest values, respectively, and it is evident that

Sequence	Frames	Image size	UAV target size (pixels)	Background details	UAV distance (m)	Focal length (mm)
Seq. 1	2966	640×512	50-90	Sky, wires	500-1000	300
Seq. 2	2087		100-120	Strong cloud, wires, buildings	200-400	75
Seq. 3	3379		80-120	Strong cloud, distractor	200-210	75
Seq. 4	2700		30-70	Strong cloud, wires, trees	300-350	75
Seq. 5	2582		100-200	Trees, telegraph pole	60-80	75

Table 1: The details of the five real test sequences for the infrared UAV target tracking, which contain species: frames, image size, UAV target size, background details, UAV distance, and focal length.

Tracking methods	P	R	$F1$	acc	FPS
TLD	0.350	0.353	0.352	0.112	12.1
SiamRPN-LT	0.701	0.708	0.705	0.470	20.5
SPLT	0.433	0.434	0.471	0.284	8.3
GlobalTrack	0.925	0.926	0.915	0.619	9.0
Ours	0.976	0.976	0.976	0.703	37.1

Table 2: Quantitative tracking results for different long-term tracking methods on the testing sequences of our infrared UAV dataset. Red and blue represent the highest value and the second highest value, respectively.

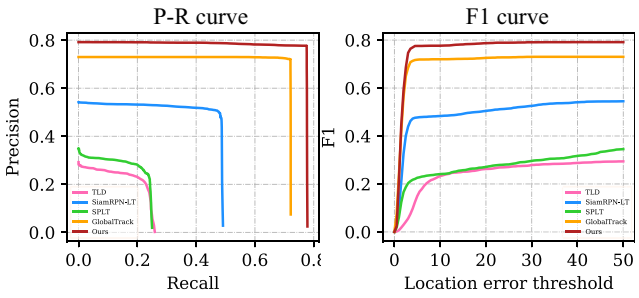

Figure 3: P-R and F1 curves of five methods. A larger area under the curve means that the method has a better tracking performance.

the proposed method surpasses others in all aspects, and that ours achieves the best scores in all metrics, where P, R, $F1$ and acc are 0.976, 0.976, 0.976 and 0.703, respectively. Our method also has the best real-time tracking performance with a tracking frames per second (FPS) of 37.1. Compared with the second best method GlobalTrack, the proposed method not only leads in tracking accuracy but also is 4.12x faster. In addition, we also give the P-R and F1 curves of five methods in Fig. 3. In general, a larger area under the P–R and F1 curves indicates that the method has a higher tracking accuracy. It can be seen that our method shows optimal performance among five methods in Fig. 3.

We analyze the cause of these results: for TLD, we believe its insufficient modality for modeling small-scaled UAV target in complex background leads to unimpressive tracking performance; for SiamRPN-LT, it mainly suffers from its backbone network AlexNet for its limited feature extraction capability; for SPLT, due to its target recapture strategy, it struggles to determine if any target exists in the input image as the background becomes complicated, and its enormous complexity jeopardizes its tracking speed; for GlobalTrack, the adoption of two-stage tracking mechanism increases the tracking precision but decreases the tracking speed, which we presume is not compatible with the requirements of practical applications. Above all, the proposed method applies a simple but effective global perception mechanism, which helps obtain an acceptable balance between tracking accuracy and speed.

The qualitative comparison is shown in Fig. 4. Each row in Fig. 4 represents typical infrared tracking result images for each sequence, where each purple box, blue box, green box, orange box, and red box are the visual descriptors of tracking result bounding boxes from TLD, SiamRPN-LT, SPLT, GlobalTrack, and the proposed method, respectively. Apparently, the proposed method obtains better tracking effect, whereas other methods are less satisfactory. We analyze the results according to the tracking difficulties as follows.

Complex background. From Seq. 2 to Seq. 5, the main backgrounds are characterized by strong cloud and trees. TLD, SiamRPN-LT and SPLT are inconsistent dealing with complex backgrounds, and thus unable to further track accurately. Although GlobalTrack employs a global perception mechanism to promptly correct false tracking, it is still easy to mis-track due to limited visual similarity calculation, *e.g.*, in Seq. 4(a), (c) and (d) GlobalTrack regards the bird and the telephone pole as the UAV target. However, benefiting from the CFRM and multi-scale structure, the proposed method is able to extract target features precisely, even when the target is extremely small as Seq. 4(c), (d) and (e).

Distrctor objects. Seq. 3(a)-(c) represent a typical distractor scenario. For SPLT, it completely failed to track throughout. As for TLD and GlobalTrack, it mis-track the

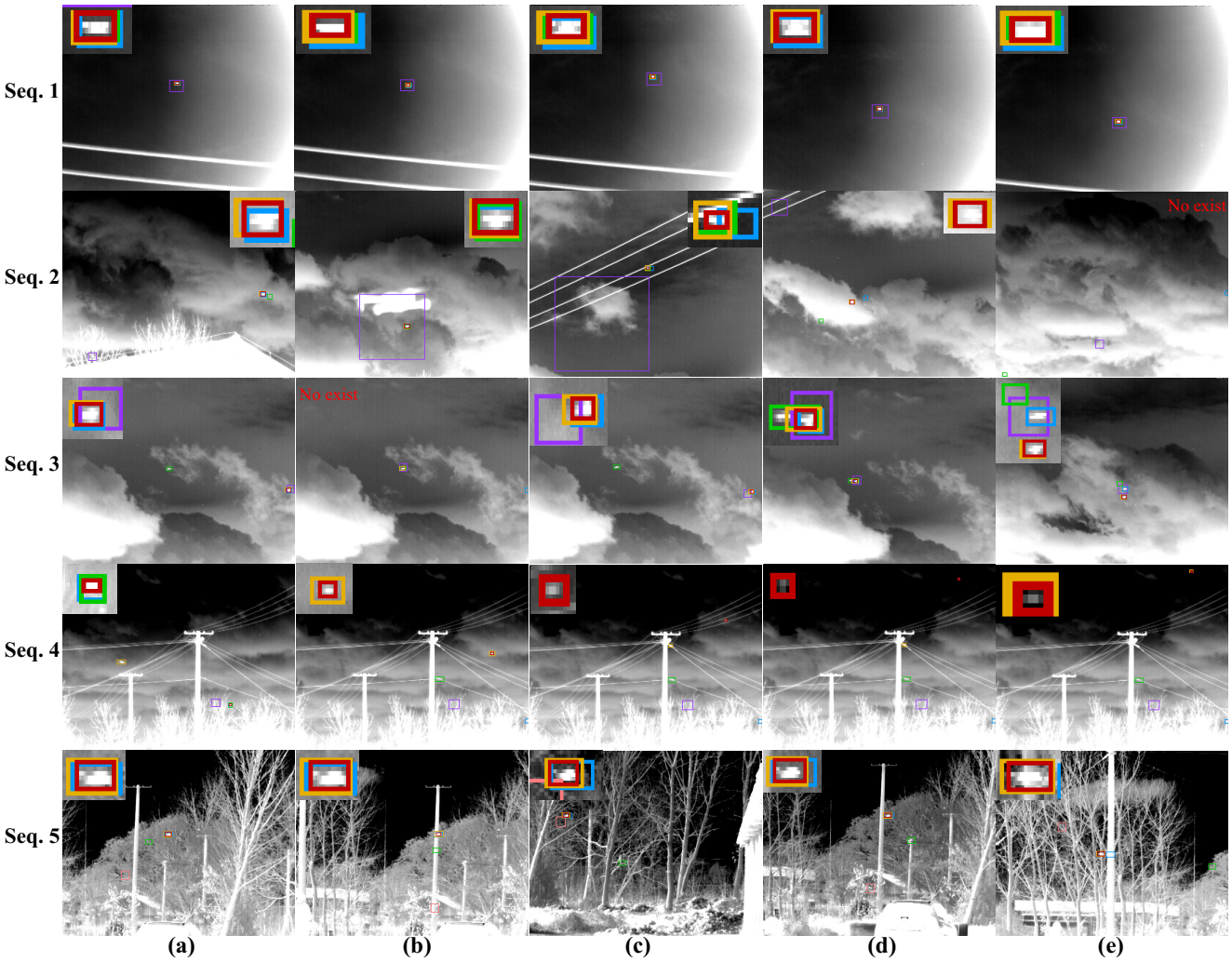

Figure 4: Qualitative tracking results. Each row displays 5 typical frames of one sequence from (a) to (e), with tracking results from different tracking methods: TLD (purple box), SiamRPN-LT (blue box), SPLT (green box), GlobalTrack (orange box), and the proposed method (red box), respectively. A method that performs better results in its corresponding bounding boxes enclosing the UAV target better. Close-ups for each tracking result are placed at the corner of each image.

distractor when the real target leaves out of view. And SiamRPN-LT believes where the target disappears still contains the target. However, the proposed method is capable of correctly discriminating between the real target and the distractor even if the real target disappears. This is attributable to the strong ability of feature extraction of the proposed method and the anti-distractor module. On one hand, it can extract target features accurately and match with the template precisely. On the other hand, it uses spatial-temporal information to exclude distrators.

Target absence. In Seq. 2(e) and Seq. 3(b), the UAV target leaves out of the field of view, and the proposed method is the only one that has successfully avoid false tracking, while none of the others yield correct results. It comes down to the global perception mechanism and the multi-scale channel attention mechanism we use, which is able to exactly perceive target absence without missed or false targets.

3.4. Ablation study

We perform ablations focusing on the impact of the CFRM and the anti-distractor module based on spatial-temporal information. We use the same amount of training and testing data and the same training criteria as discussed in Section 3.2. The results of the ablation study are given in Table 3. From the first and second rows of Table 3, we can see that all the evaluation metrics increase except acc, this is because the anti-distractor module can only eliminate the

Anti-distractor module	CFRM	Evaluation metrics			
		P	R	$F1$	acc
-	-	0.960	0.962	0.962	0.683
✓	-	0.965	0.963	0.964	0.683
-	✓	0.968	0.969	0.968	0.699
✓	✓	**0.976**	**0.976**	**0.976**	**0.703**

Table 3: Ablation studies for the proposed method. The first two columns of the table specify different versions of the proposed method.

distractor, and it is more useful for improving the performance of P, R, and $F1$; while the acc is related to the IoU between the tracked bounding box and ground-truth bounding box, so there is less improvement about acc. From the first and third rows of Table 3, we can observe that all the evaluation metrics increase due to the introduction of the CFRM module, and they improved more than that of only using the anti-distractor module in the second row. It could be properly explained by the CFRM highlighting the infrared UAV target from the complex background. As for the last row, we can see that integrating the two modules can effectively improve tracking performance. This validates the effectiveness of these two modules.

4. Conclusion

We propose a real-time anti-distractor infrared UAV tracker with global perception mechanism and channel feature refinement module for the anti-UAV mission. We combine YOLOv3 with the siamese network-based tracking framework, which is conducive to fast and long-term tracking for UAV targets. We build the channel feature refinement module through a multi-scale structure, which can further enhance the ability to extract UAV target features. And the anti-distractor module eliminates the influence of false targets from the spatial-temporal dimension and enhance the robustness of the tracking algorithm. On the five test sequences, our method achieves the best performance compared with the state-of-the-art methods, and its optimal real-time performance reaches 37.1 FPS.

5. Acknowledgements

The authors would like to thank Dr. Ming Zhang for correcting the grammatical errors.

References

[1] X. Shi, C. Yang, W. Xie, C. Liang, Z. Shi, and J. Chen, "Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 68–74, 2018.

[2] S. S. Nikolić, "An innovative response to commercial uav menace: Anti-uav falconry," *Vojno delo*, vol. 69, no. 4, pp. 146–167, 2017.

[3] R. Clothier and R. Walker, "Determination and evaluation of uav safety objectives," in *Proceedings of the 21st International Conference on Unmanned Air Vehicle Systems*, pp. 18.1–18.16, 2006.

[4] I. Guvenc, F. Koohifar, S. Singh, M. L. Sichitiu, and D. Matolak, "Detection, tracking, and interdiction for amateur drones," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 75–81, 2018.

[5] W. Liu, X. Meng, W. Qian, M. Wan, and Q. Chen, "Infrared small target detection algorithm based on multi-directional derivative and local contrast," in *AOPC 2019: Optical Sensing and Imaging Technology*, vol. 11338, pp. 1133825(1)–1133825(6), 2019.

[6] H. Fang, M. Chen, X. Liu, and S. Yao, "Infrared small target detection with total variation and reweighted l1 regularization," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–19, 2020.

[7] H. Fang, M. Xia, G. Zhou, Y. Chang, and L. Yan, "Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[8] T.-W. Bae, "Small target detection using bilateral filter and temporal cross product in infrared images," *Infrared Physics & Technology*, vol. 54, no. 5, pp. 403–411, 2011.

[9] S. Du and S. Wang, "An overview of correlation-filter-based object tracking," *IEEE Transactions on Computational Social Systems*, pp. 1–14, 2021.

[10] X. Li, Y. Zha, T. Zhang, Z. Cui, W. Zuo, Z. Hou, H. Lu, and H. Wang, "Survey of visual object tracking algorithms based on deep learning," *Journal of Image and Graphicsss*, vol. 24, no. 12, pp. 2057–2080, 2019.

[11] Y. Zhao, H. Shi, X. Chen, X. Li, and C. Wang, "An overview of object detection and tracking," in *2015 IEEE International Conference on Information and Automation*, pp. 280–286, 2015.

[12] J. F. Henriques, R. Caseiro, P. Martins, and 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.

[13] 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, 2012.

[14] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4293–4302, June 2016.

[15] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of the European Conference on Computer Vision Workshops*, pp. 850–865, 2016.

[16] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2805–2813, July 2017.

[17] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3074–3082, December 2015.

[18] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proceedings of the European Conference on Computer Vision*, vol. 9909, pp. 472–488, 2016.

[19] 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 (CVPR)*, pp. 6638–6646, July 2017.

[20] H. Fan and H. Ling, "Sanet: Structure-aware network for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 42–49, July 2017.

[21] H. Nam, M. Baek, and B. Han, "Modeling and propagating cnns in a tree structure for visual tracking," *CoRR*, vol. abs/1608.07242, 2016.

[22] R. Tao, E. Gavves, and A. W. Smeulders, "Siamese instance search for tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1420–1429, June 2016.

[23] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8971–8980, June 2018.

[24] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1328–1338, June 2019.

[25] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold siamese network for real-time object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4834–4843, June 2018.

[26] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1369–1378, June 2019.

[27] H. Fan and H. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7952–7961, June 2019.

[28] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4660–4669, June 2019.

[29] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4282–4291, June 2019.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[31] B. Yan, H. Zhao, D. Wang, H. Lu, and X. Yang, "'Skimming-Perusal' tracking: A framework for real-time and robust long-term tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2385–2393, October 2019.

[32] L. Huang, X. Zhao, and K. Huang, "Globaltrack: A simple and strong baseline for long-term tracking," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11037–11044, 2020.

[33] P. Voigtlaender, J. Luiten, P. H. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6578–6588, June 2020.

[34] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, June 2018.

[36] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[37] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, Oct 2017.

[38] A. Lukeźič, L. Č. Zajc, T. Vojíř, J. Matas, and M. Kristan, "Performance evaluation methodology for long-term single-object tracking," *IEEE transactions on cybernetics*, pp. 1–14, 2020.

[39] N. Jiang, K. Wang, X. Peng, X. Yu, Q. Wang, J. Xing, G. Li, J. Zhao, G. Guo, and Z. Han, "Anti-UAV: A large multi-modal benchmark for UAV tracking," *CoRR*, vol. abs/2101.08466, 2021.