Jointly Modeling Association and Motion Cues for Robust Infrared UAV Tracking

Boyue Xu · Ruichao Hou · Jia Bei · Tongwei Ren · Gangshan Wu

Received: date / Accepted: date

Abstract UAV tracking plays a crucial role in computer vision by enabling real-time monitoring UAVs, enhancing safety and operational capabilities while expanding the potential applications of drone technology. Off-the-shelf deep learning based trackers have not been able to effectively address challenges such as occlusion, complex motion, and background clutter for UAV objects in infrared modality. To overcome these limitations, we propose a novel tracker for UAV object tracking, named MAMC. To be specific, the proposed method first employs a data augmentation strategy to enhance the training dataset. We then introduce a candidate target association matching method to deal with the problem of interference caused by the presence of a large number of similar targets in the infrared pattern. Next, it leverages a motion estimation algorithm with window jitter compensation to address the tracking instability due to background clutter and occlusion. In addition, a simple yet effective object re-search and update strategy is used to address the complex motion and localization problem of UAV objects. Experimental results demonstrate that the proposed tracker achieves state-of-the-art performance on the Anti-UAV and LSOTB-TIR dataset.

Keywords Object tracking · UAV · infrared modality · motion estimation

1 Introduction

Unmanned aerial vehicles (UAVs) have gained significant momentum in recent years due to their ability to operate in diverse applications, such as photography [1], transportation [2] and classification [3, 4]. However, the technology has also raised significant security and privacy concerns. Therefore, to avoid certain hazards caused by threatening UAVs, it is important to detect and continuously track their locations. To address these issues, an effective tracking system for UAVs is urgently needed. Traditional visual object tracking methods [5–7] face significant challenges when tracking UAVs due to their small size and agile movements, especially in complex backgrounds or low-light conditions [8]. These challenges can be mitigated by employing infrared modalities, which are capable of capturing the thermal radiation of objects in any environment, regardless of weather conditions. HMFT [9] effectively combines RGB and infrared data for aerial tracking, resulting in impressive performance. However, the solution introduces new challenges, as UAV targets in the infrared modality only provide contour information, which can be easily confused with images generated by other thermal radiation. Moreover, tracking UAV objects in complex situations, such as fast-moving targets and objects that are lost for a long duration, remains a major challenge in infrared modalities. With the development of deep learning, most current object tracking methods can be broadly classified into two categories. The first category is based on the Siamese network [10–13], which utilizes relevance matching to identify the region that best matches the target template within a given search region. Representative methods in this category include SiamFC [10] and SiamRPN [12]. With the development of transformer [14, 15], the correlation ability between search region and templates has been further improved [5, 13]. However, these methods suffer from the common problem of selecting an appropriate search region, which is not suitable for fast moving objects and may result

All Authors are with State Key Laboratory for Novel Software Technology Nanjing University, Nanjing, China. (Email: {xuby,rc_hou}@smail.nju.edu.cn,{beijia,rentw,gswu}@nju.edu.cn). Corresponding author: Jia Bei (Email:beijia@nju.edu.cn).

in target drift. The second category of tracking methods is known as tracking by detection, exemplified by ATOM [16] and DIMP [17]. These methods first employ a detector to localize the target within the image and subsequently perform association matching to generate the tracking output. Due to the flexible motion of the UAV target, tracking by detection methods is better suited for UAV tracking tasks. The Keep-Track [18] has recently been introduced as a robust and highly effective tracking method. However, it still faces significant challenges such as occlusion and complex motion that require practical solutions.

In this paper, we propose a novel tracking network named MAMC. The proposed network consists of a detector to identify candidate objects and an object association network to filter out irrelevant objects. To overcome the challenges of UAV tracking, we propose a motion estimation module to mitigate the object tracking failure due to background clutter and occlusion. In addition, we employ a target re-search and update module to resolve issues related to complex motions. We evaluate the proposed tracker using the challenging Anti-UAV dataset [19], which includes multiple scenarios involving small objects, complex movement and occlusions, among other challenges. The proposed tracker achieves the best results on the dataset, demonstrating its efficacy in complex scenarios.

The main contributions of our work are summarized as follows:

- We propose a robust tracker for UAV tracking in infrared modality, which can effectively solve the problem of unstable UAV tracking in complex environments.
- We propose a motion estimation module to address the occlusion and background clutter problems faced by UAV tracking in the infrared modality.
- We design the target re-search and update module to address the complex motion trajectory and small target problems.

2 Related Work

2.1 UAV Tracking

UAV tracking is one of the sub-tasks of visual object tracking that has received considerable attention these years due to the challenges posed by the small size and high speed of UAV targets. To overcome these challenges, some trackers used discriminative correlation filter (DCF) [20] to track and combine spatiotemporal contextual information [21, 22]. Others attempted to improve the tracking accuracy of small UAVs by using detection-based tracking methods [23]. Additionally, attention mechanisms and siamese networks [10] were combined to accurately locate targets in images [24]. Some scholars also attempted to address the drawback of network tracking of UAVs through post-processing [25]. However, these methods suffer from some drawbacks, such as slow speed or inadequate handling of small targets, making it difficult to address the current challenges of UAV tracking.

2.2 Infrared Object Tracking

With the limitations of RGB trackers in dealing with dark nights and complex backgrounds, infrared object tracking emerged as a promising solution in recent years [26–28]. However, infrared object tracking faces the main challenge of low resolution, resulting in the availability of only contour information. Therefore, effectively extracting infrared modal information is a difficult issue. Yu *et al.* [29] proposed a method of extracting dense samples around the sample by using the Histogram of Oriented Gradients (HOG). With the development of deep learning, powerful capabilities in feature extraction were shown [30]. Recently, SiamSTA [25] and some other trackers were designed for infrared modalities and achieved excellent results. Meanwhile, HSSNet [31], also based on SiamFC, leverages a spatial awareness network to enhance discriminative ability by merging hierarchical features. However, infrared images lack color and text information, which may degrade tracking accuracy. Additionally, HSSNet used AlexNet [32] to extract features, resulting in weak feature extraction. To address the issue, MLSSNet [33] applied a spatial attention mechanism on low-level features to enhance local features and a channel attention mechanism on high-level features to distinguish features. Similarly, MMNet [34] applied a fine-grained aware network (FANet) module on low-level features to enhance fine-grained features. However, both networks used AlexNet as the backbone, which may not have a significant impact on multi-layer features.

3 Method

The structure of the method is depicted in Fig. 1. Initially, the proposed method first employs a backbone [35] to extract regional features and uses an object classifier [18] to classify them, with the highest scoring targets selected as candidates. Next, features of each candidate target are encoded along with the classifier score and target position into a vector using the feature candidate embedding network, which is referenced from the

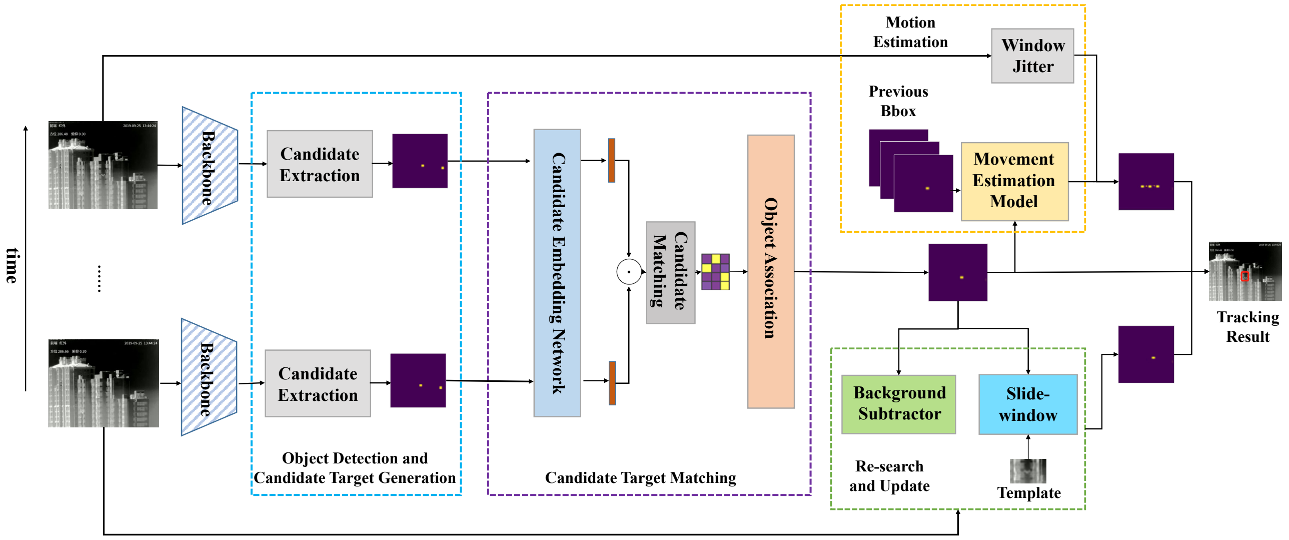

Fig. 1 The framework of the proposed method, consists of a target detection and candidate target generation module, a candidate target matching module, a motion estimation module and a target re-search and update module.

Keep-Track [18]. These feature embeddings are then used to compute a candidate feature assignment matrix between different frames. The assignment matrix represents the similarity between candidates v_i and v_j. The assignment matrix is input into the object association module [18], which associates all detected objects in the current frame with those in the previous frame and classifies them as newly appeared, kept appearing, or disappeared. The proposed method then computes the detection confidence and updates the object detection weights after association. It is the effective association operations of Keep-Track that enable it to stably track objects in interference scenarios.

The following module is the motion estimation module, which integrates motion estimation. Initially, this module calculates the background offset through motion estimation and then establishes a motion estimation prediction model to forecast the potential position of the target in the current frame. Finally, the target re-search and update module judges the target size and tracking stability. If the target is large and undergoes drift, the proposed method employs global re-search to relocate its location. In contrast, if the target is small, the module employs optical flow estimation and background separation to predict and track it. The final tracking results are obtained by considering the outcomes of all these modules together.

3.1 Motion Estimation

UAV tracking in infrared modality often encounters challenges such as occlusion or background clutter, leading to tracking failure. To address the issue, we use the method of Kalman filter [36], which is an algorithm that employs a linear equation of state to optimally estimate the system state based on the observed input and output data. It is well-suited for changing systems, especially in the field of tracking where many current target tracking algorithms, such as DeepSORT [37] and StrongSORT [38].

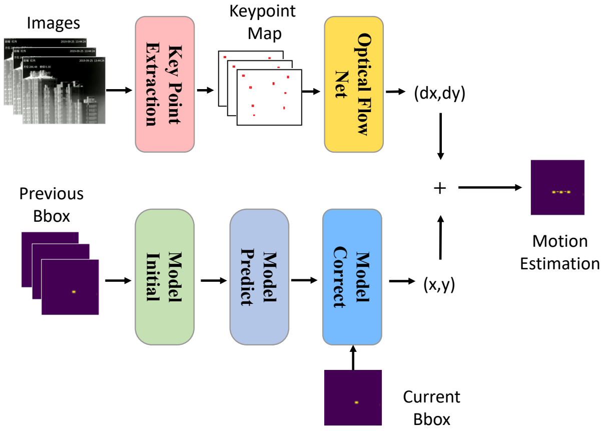

Fig. 2 Detailed design of the motion estimation module.

As shown in Fig. 2. To utilize the motion estimation for predicting small object tracking, we describe the object's trajectory in the image using four-dimensional coordinates(x, y, d_x, d_y), which respectively represent the abscissa, ordinate, lateral velocity, and longitudinal velocity of the object. We first build a motion model and an observation model of the object:

$$\mathbf{X}_k = \mathbf{A}\mathbf{X}_{k-1} + \mathbf{Q}, \tag{1}$$

$$\mathbf{Z}_k = \mathbf{H}\mathbf{X}_k + \mathbf{R}, \tag{2}$$

where $\mathbf{X}_k$ is the system state matrix, $\mathbf{Z}_k$ is the observation matrix, $\mathbf{A}$ represents the state transition matrix, which describes the position change of the object, and $\mathbf{Q}$ and $\mathbf{R}$ represent the predicted noise covariance matrix and system noise matrix, $\mathbf{H}$ represents the observation transition matrix, describing the actual motion changes of the object in each frame. To initialize the model, we use the annotation information of the first frame.

After each frame, we perform prediction and update operations. The prediction operation predicts the state of the current moment based on the system state at the previous moment as Eq.(1), and calculates the error matrix:

$$\mathbf{P}_k = \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^{\mathbf{T}} + \mathbf{Q}, \tag{3}$$

where $\mathbf{P}_k$ is the error matrix, calculated by the state transition matrix and the predicted noise covariance matrix $\mathbf{Q}$.

After the prediction, we perform an update operation that integrates the estimated state and the observed state at the current moment to estimate the optimal state:

$$\mathbf{K}_k = \mathbf{P}_k\mathbf{H}^{\mathbf{T}}\left(\mathbf{H}\mathbf{P}_k\mathbf{H}^{\mathbf{T}} + \mathbf{R}\right)^{-1}, \tag{4}$$

$$\mathbf{X}_k^{'} = \mathbf{X}_k + \mathbf{K}_k\left(\mathbf{Z}_k - \mathbf{H}\mathbf{X}_k\right), \tag{5}$$

$$\mathbf{P}_k^{'} = \mathbf{P}_k - \mathbf{K}_k\mathbf{H}\mathbf{P}_k, \tag{6}$$

where $\mathbf{K}_k$ represents the motion estimation gain, $\mathbf{R}$ represents the measurement noise covariance matrix, $\mathbf{H}$ represents the observation matrix, and $\mathbf{Z}_k$ represents the observed value at time k; $\mathbf{X}_k^{'}$ and $\mathbf{P}_k^{'}$ represent the updated system state matrix and the updated error matrix respectively. The final output of the corrected prediction state is the final prediction coordinates. To address the issue of camera shake during tracking, which causes the prediction result to shake, we calculate the window jitter of the front and rear frames using the Lucas-Kanade method [39]. We then use the offset to further correct the prediction:

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = \mathbf{D}\begin{bmatrix} -\sum_i I_x\left(q_i\right) I_t\left(q_i\right) \\ -\sum_i I_y\left(q_i\right) I_t\left(q_i\right) \end{bmatrix}, \tag{7}$$

$$\mathbf{D} = \begin{bmatrix} \sum_i I_x\left(q_i\right)^2 & \sum_i I_x\left(q_i\right) I_y\left(q_i\right) \\ \sum_i I_y\left(q_i\right) I_x\left(q_i\right) & \sum_i I_y\left(q_i\right)^2 \end{bmatrix}^{-1}, \tag{8}$$

where V_k and V_y represent the camera offset velocity horizontally and vertically at this frame, $I_x(q_i)$ and $I_y(q_i)$ represents the partial derivative of point q_i to x and y. Finally, the lensing shake offset is superimposed on the motion estimation model results to obtain the final result.

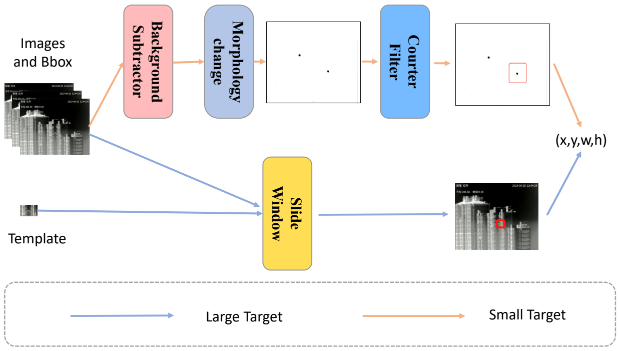

Fig. 3 Detailed design of the target re-search and update module.

3.2 Re-search and Update

UAV tracking often faces the challenge of object drift and complex motion, which can result in tracking failure. To address the issue, we propose a target re-search and update module which is shown in Fig. 3. This module mainly addresses the problem of target loss caused by complex motion trajectories, using different methods to handle targets of varying sizes.

For small targets, we use background subtraction to locate moving targets in the frame and select new candidates by computing the distance to the previously lost target. For larger targets, we use a global matching approach to directly find the lost targets. When the tracking target is lost, we first judge the size of the target. If the target is smaller than $W \times H$ pixels, we consider it to be a small target, which is easily submerged in background noise and ignored by the network. The reason for distinguishing between large and smalle targets is that UAV targets mostly have a relatively fixed shape, often appearing as rectangles in images. Measuring their width and height provides a better way to differentiate between target sizes. However, since the target is still moving relative to the background, we perform background differences using previous frames, followed by erosion and dilation operations to enlarge the moving pixels and make them more visible. We then judge the edge contours of the moving pixels, and if the contour is larger than N pixels, we consider it a candidate target. If no candidate targets are found, this module will retain the results from the previous frame. In fact, such cases are rare, as the method used by the module is able to keenly capture moving targets within the frame. If the target remains stationary, retaining the results from the previous frame can provide stable tracking of the target. Finally, we perform candidate target matching operations. We select candidate points whose distance from the nearest credible result with judgment confidence greater than K before target drifting is less than M as the current frame result. If there is no such

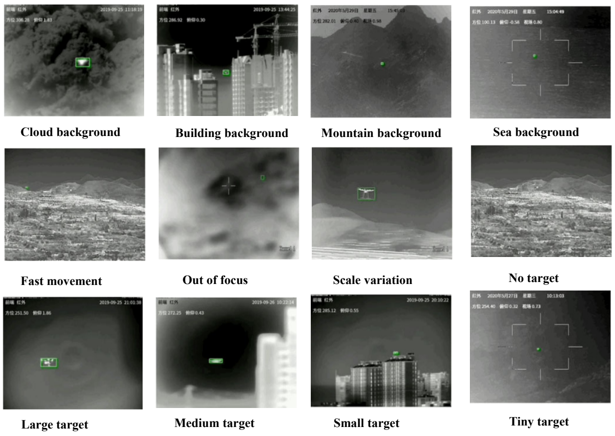

Fig. 4 Samples of Anti-UAV dataset, the ground truth is marked in green.

result, we expand the search region and try the search again.

For large targets with a size greater than $H \times W$, we employ a simple yet efficient template-matching strategy after the target is lost. We use sliding window matching with the target given in the first frame as the template and perform normalized square difference matching in the frames where the object is lost. such as the formula:

$$R(x,y)=\frac{\sum_{x',y'}\left(M\left(x',y'\right)-I\left(x+x',y+y'\right)\right)^2}{\sqrt{\sum_{z',y'}M\left(x',y'\right)^2\cdot\sum_{z',y'}I\left(x+x',y+y'\right)^2}}, \tag{9}$$

where $R(x,y)$ represents the matching degree between the current region and the template, where the closer to 0, the higher the matching degree; $M(x,y)$ and $I(x,y)$ represent the pixel in the template and search region, respectively.

We take the difference, square and sum operations to obtain the matching degree of the current location and identify the location with the highest matching degree in the image as the target location. Simultaneously, we update the network data to ensure follow-up tracking effects.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate the effectiveness of the proposed method by using the Anti-UAV Competition dataset. As shown in Fig. 4, the dataset comprises over 100 complex scenes involving UAVs of varying sizes (large, medium, small, and tiny), as well as diverse backgrounds, including clouds, mountains, buildings, and seascapes. Furthermore, the dataset presents complex situations such as occlusion, rapid motion, and targets out of sight. Most of the sequences in the dataset are long-term, providing an opportunity to test the robustness and effectiveness of the tracker in challenging scenarios. The evaluation aims to determine the reliability of the proposed method in accurately tracking UAVs in these complex scenarios. The performance of our approach is analyzed in terms of its ability to handle the aforementioned challenges.

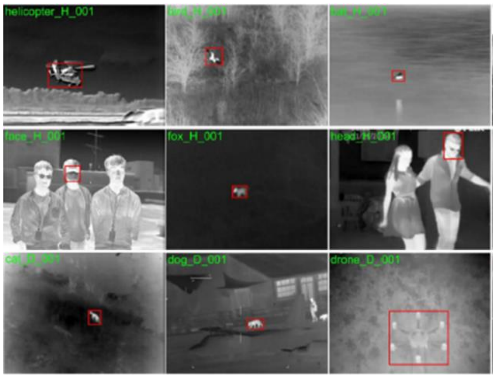

Fig. 5 Samples of LSOTB-TIR dataset, the ground truth is marked in red

.

To further validate the effectiveness of the proposed method, we also conduct experiments on the LSOTB-TIR [40] dataset. This dataset not only includes UAV targets but also contains a wide range of common infrared targets, as depicted in Figure 5.

Evaluation metric. The Anti-UAV competition employs an official evaluation metric that calculates the average Intersection over Union (IoU) of all video sequences. The accuracy is calculated as:

$$ACC=\frac{1}{T}\sum_{t=1}^{T}IoU_t\times\delta\left(v_t>0\right)+p_t\times\left(1-\delta\left(v_t>0\right)\right), \tag{10}$$

where T denotes the number of frames in the video sequence, iou_t represents the IoU of the t-th frame, p_t denotes the predicted visibility flag. Specifically, p_t equals 1 when the predicted bounding box is empty, and 0 otherwise. Moreover, v_t denotes the ground-truth visibility flag of the target. The indicator function $\delta(v_t > 0)$ is equal to 1 when $v_t > 0$, and 0 otherwise. The metric is designed to assess the performance of object detection and tracking algorithms in detecting and tracking UAVs. It provides a rigorous and objective measure of the accuracy of the predicted bounding boxes compared to the ground-truth annotations. The evaluation metric promotes transparency and fairness in the competition and enables meaningful comparisons of the performance of different algorithms.

On the LSOTB-TIR [40] dataset, we use the precision and success metrics as provided by the dataset creators. Precision refers to the accuracy of the predicted bounding box, measured as the Euclidean distance between the center points of the predicted box and the ground truth box. Generally, if the Euclidean distance between the center points of the ground truth box and the predicted box is less than 20 pixels, it is considered as accurate localization. Success rate measures the IoU between the predicted box and the ground truth box. If the IoU exceeds a certain threshold, the tracking is considered successful.

Network parameters. The proposed method is implemented using the PyTorch 1.10 platform with an i9 CPU, 64GB RAM, and a RTX3090 GPU with 24GB memory. We use the Keep-Track [18] as our baseline, and we follow its basic hyperparameters. For pre-training, we utilize the weights of baseline, while for finetuning, we employ the Anti-UAV training set [19]. To adapt to the characteristics of UAVs in infrared mode, we perform data enhancement on the dataset, including random occlusion of the target, rotation from -45° to 45°, and blurring of the target, among others. This is done to improve the generalization ability of the model to complex situations.

Furthermore, we confirm that tracking confidence greater than 0.8 indicates stable tracking results and can be directly used. In turn, tracking confidence less than 0.3 indicates tracking failure, which needs to be handled by the following two modules.

4.2 Quantitative Evaluation

The evaluation of the proposed tracker is conducted by comparing its performance with that of other competing methods on the test set of the Anti-UAV dataset [19] and LSOTB-TIR dateser [40]. To ensure fairness and reliability, all RGB-based tracking methods are fine-tuned on the training set of the corresponding infrared dataset.

The comparison results are presented in Table 1, which employs official evaluation metrics on the test set. The proposed tracker has demonstrated superior performance compared to all other trackers on the test set, achieving a score of 65.02%. The score is 0.58% higher than that of the second-ranked tracker which is the champion algorithm of the Anti-UAV competition, which introduce a three-stage re-detection mechanism to re-detect targets. Furthermore, the score is 5.94% higher than that of the Stark [41] which is one of the newest trackers. These results provide strong evidence for the effectiveness of the proposed tracker in complex drone tracking scenarios. It is worth mentioning that the PVT++ [42] is the latest motion-estimation based method, achieving the fastest tracking speed. However, our method still leads it by 10.04%. This is because the motion trajectory of UAV targets is a combination of the UAV's trajectory and the camera's trajectory, often resulting in back-and-forth movements in the image. This poses a significant challenge for accurate position estimation. The proposed motion estimation module in this method estimates not only the target's motion trajectory but also compensates for the offset caused by camera motion. The combined result provides the true trajectory of the UAV in the frame, effectively solving this problem.

Table 1 Comparison results of the proposed method against the competing trackers on Anti-UAV test set. The best results are highlighted in bold.

Method	Source	Score	FPS
HiFT [43]	ICCV21	37.87	127
SiamTPN [44]	WACV22	40.46	80
STMTrack [45]	CVPR21	40.86	37
TCTrack [46]	CVPR22	41.59	160
UDAT [47]	CVPR22	44.17	80
PVT++ [42]	ICCV23	44.98	**200**
OStrack [48]	ECCV22	46.88	60
TransT [13]	CVPR21	52.14	50
TOMP [49]	CVPR22	53.42	25
TransformerTrack [50]	CVPR21	54.75	26
RTS [51]	ECCV22	55.12	30
Stark [41]	ICCV21	59.08	26
3rd tracker	Anti-UAV 2021	63.80	-
2nd tracker	Anti-UAV 2021	63.88	-
1st tracker (winner)	Anti-UAV 2021	64.44	-
MAMC		**65.02**	35

Table 2 Comparison results of the proposed method against the competing trackers on LSOTB-TIR test set. The best results are highlighted in bold.

Method	Year	Precision	Success
HSSNet [52]	2019	0.52	0.41
MCFTS [53]	2017	0.64	0.48
SiamSAV [54]	2021	0.70	0.
58STAMT [55]	2022	0.71	0.
58GFSNet [56]	2021	0.76	0.
CMD-DiMP [57]	2021	**0.81**	0.62
MAMC		**0.81**	**0.69**

In the experiments on the LSOTB-TIR [40] test set, we compare our method with several state-of-the-art methods, as shown in Table 2. In terms of Precision, we achieve the same score as the best-performing method. However, in terms of Success, our method outperform the best-performing method by 7%.

The superior performance of the proposed tracker can be attributed to its ability to accurately localize UAVs against small targets and jamming backgrounds and self-correct after tracking problems. The ability to accurately localize UAVs in challenging scenarios is of paramount importance for real-world applications such as border control, public safety, and infrastructure protection.

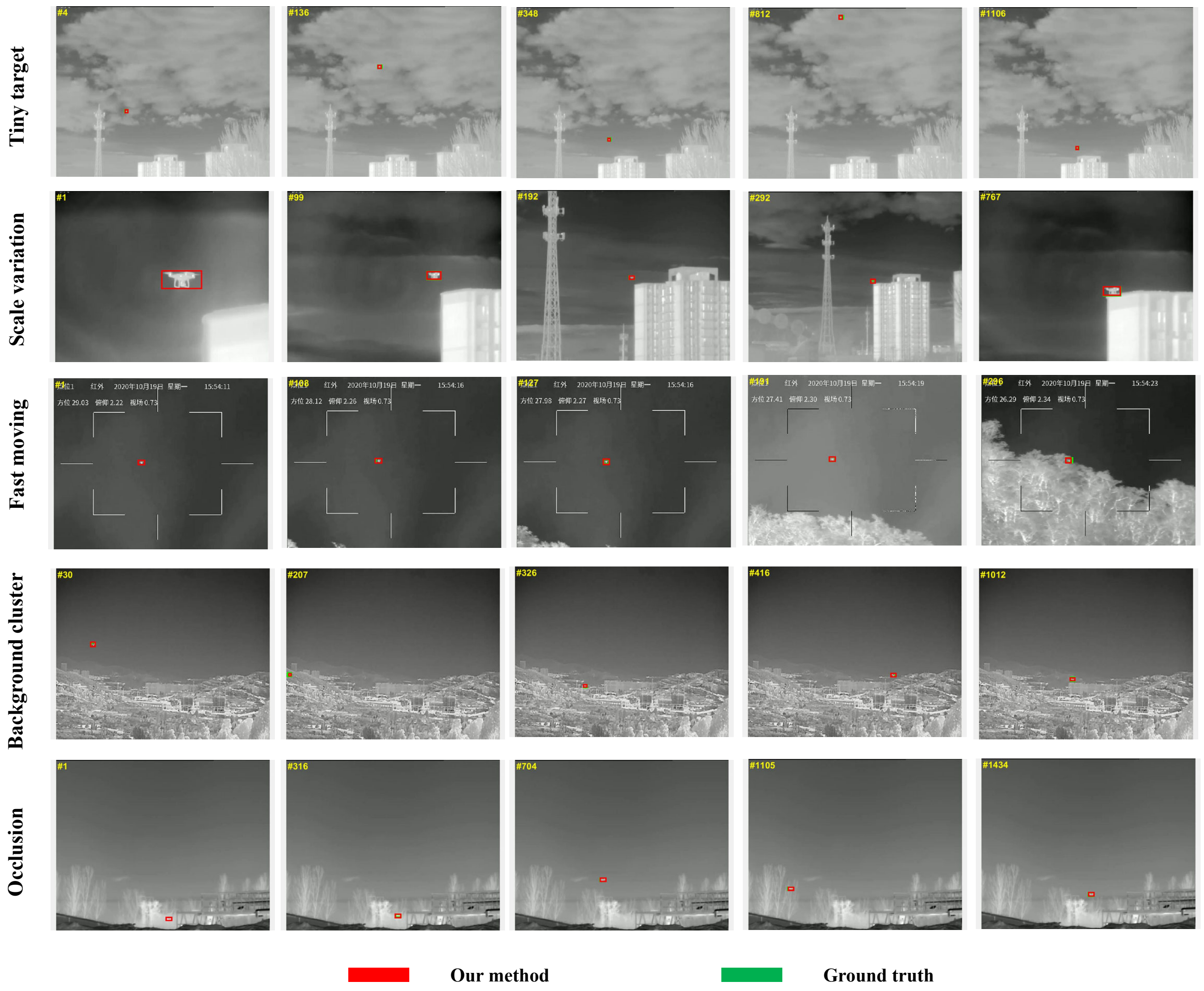

Fig. 6 Qualitative comparison of the proposed tracker in handling different challenging scenarios.

4.3 Qualitative Evaluation

In this section, we evaluate the performance of our proposed method in a variety of complex scenarios, including small objects, scale variation, fast motion, occlusion, and background clutter. The results of the proposed method are presented in Fig. 6, which illustrates its ability to track objects effectively and stably in challenging scenarios.

The success of the proposed method in tracking small targets and background-cluttered scenes can be attributed to the use of a small target re-search algorithm and motion estimation algorithm, which can identify objects for which deep learning does not work. Moreover, the proposed method demonstrates robustness and effectiveness in dealing with object drift, especially in occlusion or when the object is out of sight.

Overall, the results obtained in our work highlight the potential of the proposed method in addressing the challenges of drone tracking in complicated scenarios. The combination of a target re-search algorithm, motion estimation algorithm, and deep network-based approaches has enabled the proposed method to achieve high levels of accuracy and stability in tracking small objects, even in challenging situations.

Table 3 Ablation study on different components.

Augmentation	Motion estimation	Re-search	Score
			56.37
	✓		57.51
	✓	✓	61.34
✓			60.35
✓	✓		61.60
✓	✓	✓	**65.02**

4.4 Ablation Study

In order to assess the individual effectiveness of each module in the proposed method, we conduct an ablation experiment using the test set of the Anti-UAV competition [19]. The evaluation metric employed in the experiment remained consistent with the official metric of the competition. The baseline method used for comparison is the base Keep-Track [18]. The results of the ablation experiment are presented in Table 3. This evaluation approach allows us to investigate the impact of each module on the overall performance of the proposed method, thereby providing a more thorough understanding of the efficacy of the proposed method.

Data augmentation. To improve the ability of the proposed model to accurately track UAV objects in infrared modality, we retrain the proposed model using the training set of the Anti-UAV Competition. In addition, we employ data augmentation techniques to further improve the model performance. The resulting model achieves a 3.98% improvement in performance compared to the baseline method. These results demonstrate that such a simple operational adjustment can have a significant impact on the generalization ability of the model.

Motion estimation. Building on the data augmentation techniques used in previous experiments, we add a motion estimation module to the proposed method to effectively address occlusion and background clutter issues. The results in Table 2 demonstrate that the incorporation of this module improves the performance of the method by 1.25% compared to the data-augmentation baseline, while the version without data augmentation exhibited a 1.14% improvement compared to the baseline without data augmentation. To provide further insight into the impact of this module, Fig. 7 illustrates a scenario in which the object and the background are difficult to distinguish, and the network is faced with a challenging tracking task. When the module is not utilized, the tracking performance deteriorates significantly. However, when the motion estimation module is integrated into the method, stable tracking is maintained despite challenging scenarios. These results highlight the efficacy of the motion estimation module in improving the tracking performance of the proposed method in complex tracking scenarios.

Re-search and update. We introduce a target research module to further enhance the tracking performance. The experiments demonstrate that the network with the inclusion of this module experienced an improvement of 3.42% compared to when the module was not included. Additionally, the network that incorporated this module but did not include data augmentation also showed an improvement of 3.83% compared to the network that did not incorporate this module or data augmentation. Notably, this module proved particularly effective in complex scenarios where the target is small and the movement is complex, as demonstrated in Fig. 8. In such scenarios, the target re-search module successfully detect and tracks the moving target, resulting in stable and accurate tracking.

Table 4 Comparison of different thresholds.

	N=10	N=20	N=30
$W \times H(9\times7)$	64.24	62.73	62.97
$W\times\ H(20\times15)$	**65.02**	62.86	63.11
$W\times\ H(30\times20)$	62.45	61.82	62.09

4.5 Parameter Analysis

In this section, we perform a parameter analysis experiment, which is mainly divided into two parts. The first part focuses on the parameter analysis of the differentiation threshold for large and small targets, while the second part investigates the enhancement of tracking performance by our method on the two classes of targets after differentiating between large and small ones.

The parameter analysis for large and small targets primarily encompasses two sets of parameters namely the determination range $W\times H$ for small targets and the candidate target contour threshold N. The specific experimental results are shown in Table 4. Since most drone targets are rectangular, W and H are set to the proportions of a rectangle, and N is taken as 10, 20, 30; Targets with a perimeter larger than N threshold are considered as candidate targets, while those below are considered as noise. The experiments show that when $W\times H$ is set to 20×15, and N is 10, tracking achieves the best effect. This is because if W and H are set too small, some small targets will be misclassified as large targets, and these targets are difficult to feature extract through deep neural networks and cannot be relocated through global matching for large targets, leading to tracking failure. Conversely, setting W and H too large can misclassify some originally stable large targets as small targets, thereby disturbing the originally stable tracking.

On the basis of differentiating between large and small targets, we conduct experiments to enhance the tracking performance of both types of targets, as shown in Table 5. Experimental results demonstrate significant improvements in tracking small targets using our proposed method. With the combined effects of the motion estimation and target re-search and update modules, the tracking effectiveness for small targets was increased by 10.51%.

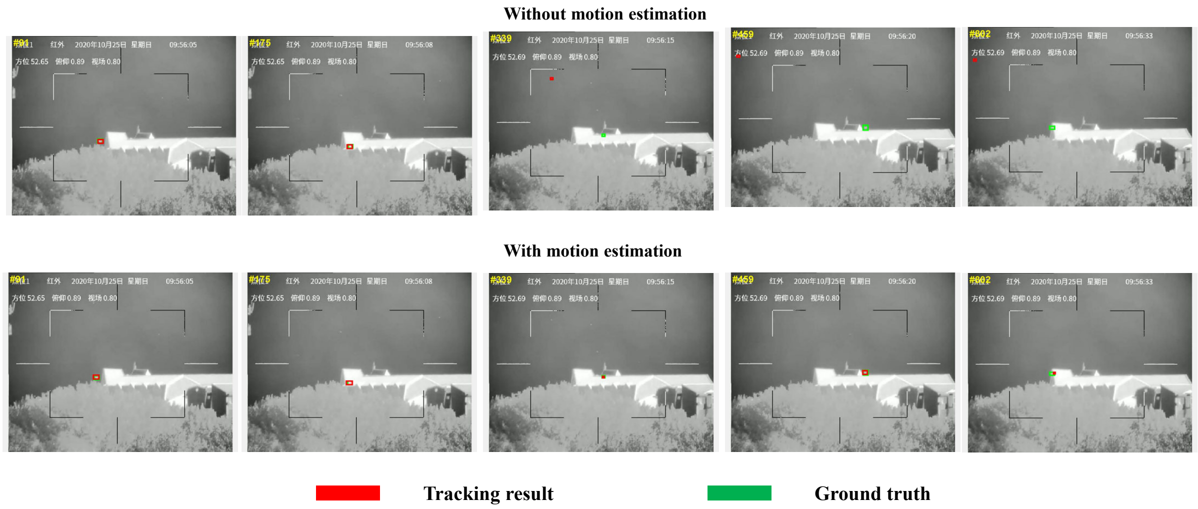

Fig. 7 Ablation study of motion estimation module.

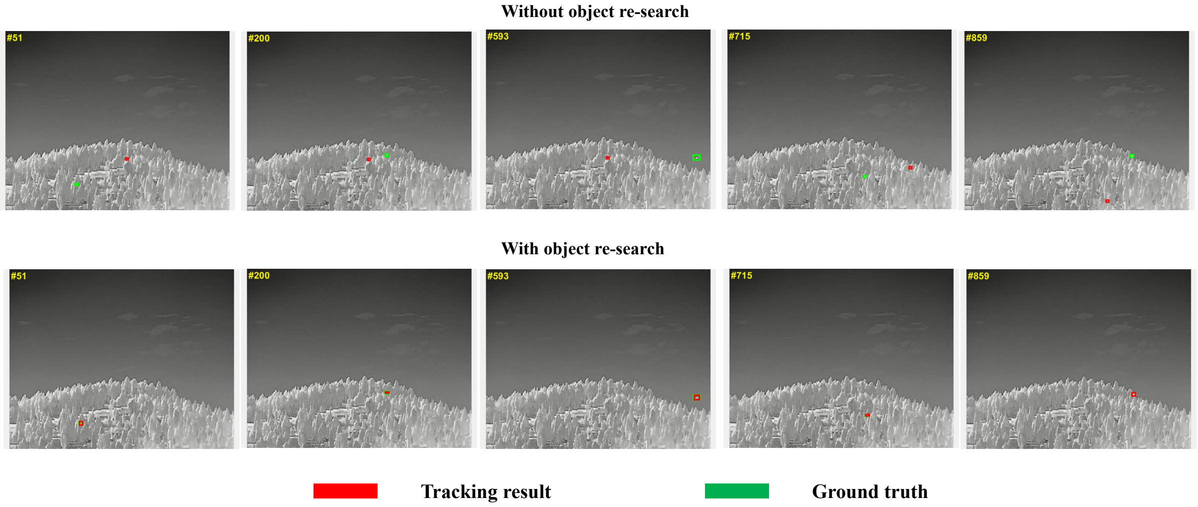

Fig. 8 Ablation study of object re-search and update module.

Table 5 Comparison of tracking performance on different targets.

	Large	Small	All
baseline	60.39	44.07	56.37
MAMC	68.38	55.48	65.02

5 Conclusion

In this paper, we proposed a novel infrared object tracker named MAMC, which effectively addresses various complex issues encountered in small object tracking. First, we augmented the dataset using data augmentation techniques to improve model generalization. We then introduced a motion estimation module to deal with occlusion and background clutter, and an object re-search module to deal with complex motions and object drifts. We validated the effectiveness of the proposed algorithm on the Anti-UAV and LSOTB-TIR dataset, achieving state-of-the-art performance. While our work has achieved good results in the current test set, drone tracking is a task of great practical value, and working solely on limited datasets is insufficient.

Declarations

Funding

This work was supported by the National Natural Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (021714380026), the program B for Outstanding Ph.D, candidate of Nanjing University, and the Collaborative

Innovation Center of Novel Software Technology and Industrialization.

Author contributions

Boyue Xu: Conceptualization, Methodology, Software, Writing – original draft. Ruichao Hou: Investigation, Methodology, Validation. Jia Bei: Project administration, Writing – original draft. Tongwei Ren: Polishing, Funding acquisition, Recource. Gangshan Wu: Supervision, Funding acquisition.

References

1. Nan Jiang, Bin Sheng, Ping Li, and Tong-Yee Lee. Photohelper: Portrait photographing guidance via deep feature retrieval and fusion. *IEEE Transactions on Multimedia*, 2022.
2. Zhihua Chen, Jun Qiu, Bin Sheng, Ping Li, and Enhua Wu. Gpsd: generative parking spot detection using multi-clue recovery model. *The Visual Computer*, 37(9-11):2657–2669, 2021.
3. Abdulrhman H Al-Jebrni, Saba Ghazanfar Ali, Huating Li, Xiao Lin, Ping Li, Younhyun Jung, Jinman Kim, David Dagan Feng, Bin Sheng, Lixin Jiang, et al. Sthy-net: a feature fusion-enhanced dense-branched modules network for small thyroid nodule classification from ultrasound images. *The Visual Computer*, pages 1–15, 2023.
4. Jiajia Li, Jie Chen, Bin Sheng, Ping Li, Po Yang, David Dagan Feng, and Jun Qi. Automatic detection and classification system of domestic waste via multimodel cascaded convolutional neural network. *IEEE transactions on industrial informatics*, 18(1):163–173, 2021.
5. Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
6. Rudrika Kalsotra and Sakshi Arora. Background subtraction for moving object detection: explorations of recent developments and challenges. *The Visual Computer*, 38(12):4151–4178, 2022.
7. Mohammed Y Abbass, Ki-Chul Kwon, Nam Kim, Safey A Abdelwahab, Fathi E Abd El-Samie, and Ashraf AM Khalaf. A survey on online learning for visual tracking. *The Visual Computer*, 37:993–1014, 2021.
8. Yabin Zhu, Chenglong Li, Yao Liu, Xiao Wang, Jin Tang, Bin Luo, and Zhixiang Huang. Tiny object tracking: A large-scale dataset and a baseline. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2023.
9. Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022.
10. Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision Workshops*, 2016.
11. Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI Conference on Artificial Intelligence*, 2020.
12. Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
13. X. Chen, B. Yan, J. Zhu, et al. Transformer tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
14. Xiao Lin, Shuzhou Sun, Wei Huang, Bin Sheng, Ping Li, and David Dagan Feng. Eapt: efficient attention pyramid transformer for image processing. *IEEE Transactions on Multimedia*, 2021.
15. Zhifeng Xie, Wenling Zhang, Bin Sheng, Ping Li, and CL Philip Chen. Bagfn: broad attentive graph fusion network for high-order feature interactions. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
16. Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
17. Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *IEEE/CVF International Conference on Computer Vision*, 2019.
18. Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *IEEE/CVF International Conference on Computer Vision*, 2021.
19. Jian Zhao, Gang Wang, Jianan Li, Lei Jin, Nana Fan, Min Wang, Xiaojuan Wang, Ting Yong, Yafeng Deng, Yandong Guo, et al. The 2nd anti-uav workshop & challenge: methods and results. *arXiv preprint arXiv:2108.09909*, 2021.
20. Jianming Zhang, Tingyu Yuan, Yaoqi He, and Jin Wang. A background-aware correlation filter with adaptive saliency-aware regularization for visual tracking. *Neural Computing and Applications*, 2022.
21. Di Yuan, Xiaojun Chang, Zhihui Li, and Zhenyu He. Learning adaptive spatial-temporal context-aware correlation filters for uav tracking. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(3):1–18, 2022.
22. Jiwei Fan, Xiaogang Yang, Ruitao Lu, Weipeng Li, and Yueping Huang. Long-term visual tracking algorithm for uavs based on kernel correlation filtering and surf features. *The Visual Computer*, 39(1):319–333, 2023.
23. Jie Zhao, Jingshu Zhang, Dongdong Li, and Dong Wang. Vision-based anti-uav detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):25323–25334, 2022.
24. Xiaoran Shi, Yan Zhang, Zhiguang Shi, and Yu Zhang. Gasiam: Graph attention based siamese tracker for infrared anti-uav. In *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications*, 2022.
25. Bo Huang, Junjie Chen, Tingfa Xu, Ying Wang, Shenwang Jiang, Yuncheng Wang, Lei Wang, and Jianan Li. Siamsta: Spatio-temporal attention based siamese tracker for tracking uavs. In *IEEE/CVF International Conference on Computer Vision*, 2021.
26. Ruichao Hou, Tongwei Ren, and Gangshan Wu. Mirnet: A robust rgbt tracking jointly with multi-modal interaction and refinement. In *IEEE International Conference on Multimedia and Expo*, 2022.

27. Ruichao Hou, Boyue Xu, Tongwei Ren, and Gangshan Wu. Mtnet: Learning modality-aware representation with transformer for rgbt tracking. In *IEEE International Conference on Multimedia and Expo*, 2023.
28. Andong Lu, Cun Qian, Chenglong Li, Jin Tang, and Liang Wang. Duality-gated mutual condition network for rgbt tracking. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022.
29. Xianguo Yu and Qifeng Yu. Online structural learning with dense samples and a weighting kernel. *Pattern Recognition Letters*, 105:59–66, 2018.
30. Han Wu, Weiqiang Li, Wanqi Li, and Guizhong Liu. A real-time robust approach for tracking uavs in infrared videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.
31. Qiao Liu, Xiaohuan Lu, Zhenyu He, Chunkai Zhang, and Wen-Sheng Chen. Deep convolutional neural networks for thermal infrared object tracking. *Knowledge-Based Systems*, 134:189–198, 2017.
32. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing systems*, 25, 2012.
33. Qiao Liu, Xin Li, Zhenyu He, Nana Fan, Di Yuan, and Hongpeng Wang. Learning deep multi-level similarity for thermal infrared object tracking. *IEEE Transactions on Multimedia*, 23:2114–2126, 2020.
34. Qiao Liu, Di Yuan, Nana Fan, Peng Gao, Xin Li, and Zhenyu He. Learning dual-level deep representation for thermal infrared tracking. *IEEE Transactions on Multimedia*, 25:1269–1281, 2022.
35. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
36. Gregory F Welch. Kalman filter. *Computer Vision: A Reference Guide*, 1:1–3, 2020.
37. Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*, 2017.
38. Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023.
39. Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2011.
40. Qiao Liu, Xin Li, Zhenyu He, Chenglong Li, Jun Li, Zikun Zhou, Di Yuan, Jing Li, Kai Yang, Nana Fan, et al. Lsotb-tir: A large-scale high-diversity thermal infrared object tracking benchmark. In *Proceedings of the 28th ACM international conference on multimedia*, 2020.
41. Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *IEEE/CVF International Conference on Computer Vision*, 2021.
42. Bowen Li, Ziyuan Huang, Junjie Ye, Yiming Li, Sebastian Scherer, Hang Zhao, and Changhong Fu. Pvt++: A simple end-to-end latency-aware visual tracking framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
43. Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Hift: Hierarchical feature transformer for aerial tracking. In *IEEE/CVF International Conference on Computer Vision*, 2021.
44. Daitao Xing, Nikolaos Evangeliou, Athanasios Tsoukalas, and Anthony Tzes. Siamese transformer pyramid networks for real-time uav tracking. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.
45. Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
46. Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. Tctrack: temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
47. Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
48. Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, 2022.
49. Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
50. Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
51. Matthieu Paul, Martin Danelljan, Christoph Mayer, and Luc Van Gool. Robust visual tracking by segmentation. In *European Conference on Computer Vision*, 2022.
52. Xin Li, Qiao Liu, Nana Fan, Zhenyu He, and Hongzhi Wang. Hierarchical spatial-aware siamese network for thermal infrared object tracking. *Knowledge-Based Systems*, 166:71–81, 2019.
53. Qiao Liu, Xiaohuan Lu, Zhenyu He, Chunkai Zhang, and Wen-Sheng Chen. Deep convolutional neural networks for thermal infrared object tracking. *Knowledge-Based Systems*, 134:189–198, 2017.
54. Tingting Yao, Jincheng Hu, Bo Zhang, Yuan Gao, Pengfei Li, and Qing Hu. Scale and appearance variation enhanced siamese network for thermal infrared target tracking. *Infrared Physics & Technology*, 117:103825, 2021.
55. Di Yuan, Xiu Shu, Qiao Liu, and Zhenyu He. Structural target-aware model for thermal infrared tracking. *Neurocomputing*, 491:44–56, 2022.
56. Ruimin Chen, Shijian Liu, Zhuang Miao, and Fanming Li. Gfsnet: generalization-friendly siamese network for thermal infrared object tracking. *Infrared Physics & Technology*, 123:104190, 2022.
57. Jingxian Sun, Lichao Zhang, Yufei Zha, Abel Gonzalez-Garcia, Peng Zhang, Wei Huang, and Yanning Zhang. Unsupervised cross-modal distillation for thermal infrared tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.