A Global-Local Tracking Framework Driven by Both Motion and Appearance for Infrared Anti-UAV

Yifan Li
Xidian University
Xi'an, China
18066899461@163.com

Dian Yuan
Xidian University
Xi'an, China
d1anskr@163.com

Meng Sun
Xidian University
Xi'an, China
sun_meng2002@163.com

Hongyu Wang
Xidian University
Xi'an, China
22171214782@stu.xidian.edu.cn

Xiaotao Liu
Xidian University
Xi'an, China
xtliu@xidian.edu.cn

Jing Liu
Xidian University
Xi'an, China
neouma@163.com

Abstract

*Unmanned aerial vehicles (UAVs) have been widely used in various application domains, but unauthorized UAVs may pose a threat to public safety due to violation of aviation regulations. Therefore, how to design an effective UAV tracking method for anti-UAV is a crucial part of the UAV-defense system. In this paper, we propose a Global-Local Tracking Framework driven by both Motion and Appearance (**GLTF-MA**) including four modules to deal with the practical difficulties in infrared anti-UAV. Firstly, a Periodic Global Detection (PGD) module is periodically performed to re-locate UAVs in the whole image to account for frequent appearance/disappearance and unstable flight paths of UAVs. Meanwhile, a Multi-stage Local Tracking (MLT) module containing a priori stage switching mechanism, motion-appearance matching mechanism, and a motion estimation punisher is routinely implemented to deal with the tiny size of UAVs and background interference. Next, a Target Disappearance Judgement (TDJ) module is performed to give a robust target disappearance flag, followed by a Bounding Box Refinement (BBR) module to refine the target box when the TDJ module thinks the target exists. Extensive experiments demonstrate the superiority of GLTF-MA over other competing counterparts, especially when the UAV is low resolution and moves quickly.*

1. Introduction

In recent years, unmanned aerial vehicles (UAVs) have been widely used in wildlife monitoring, crowd monitoring/management, and videography of extreme sports [1], etc. Nevertheless, unauthorized UAVs may violate aviation safety regulations, thereby posing a potential threat to public safety such as airport disruptions and flight delays. Nowadays, it is highly desired to develop anti-UAV techniques to defend against these UAV accidents.

Common sensors in the traditional anti-UAV systems consist of radar [2], radio [3], and sonar [4], yet their performance is suboptimal because relatively small UAVs are difficult to perceive by these sensors. Recently, infrared imaging technology has been prevalent for UAV tracking with its all-day imaging capability and stability in harsh conditions [5]. However, there are several difficulties in infrared anti-UAV: (i) UAV frequently disappearing/appearing in view due to low flight altitudes and being occluded, (ii) erratic flight paths, (iii) tiny size and lack of appearance information, (iv) background interference in complex environments such as forests and buildings. Hence, it is still a challenging task to detect and track UAVs in infrared videos with high accuracy. [6, 7]

With the rapid development of computer hardware, deep learning-based trackers [8–12] play a dominant role in the field of target tracking. The framework of these trackers typically contains three main components: A backbone to extract deep features of the templates and the search region, an integration module to fuse the features of the templates and the search region, and a prediction head to locate the target. While considering tracking UAVs, which contains a wide range of target occlusion and erratic flight paths situations, it will degrade the performance of such a general tracking framework that only locates the target within a local search region. To alleviate this problem, we design a global-local tracking frame to switch between the local search region and the entire image for robust tracking. In addition, infrared UAVs are characterized by low resolution

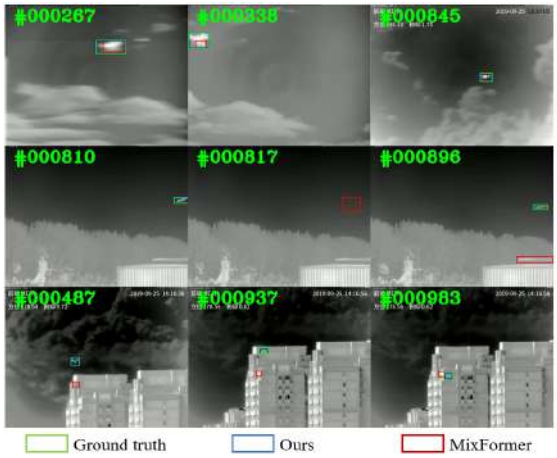


Figure 1. Visualization results of GLTF-MA (ours) and MixFormer on the 3rd Anti-UAV challenge. Compared with MixFormer, in the first row, the prediction box of GLTF-MA is more accurate; In the second row, GLTF-MA can find the target faster when the target reappears in the view; In the third row, GLTF-MA can better resist background interference.

and lack of appearance information, which is prone to make the background objects mislead these trackers without target temporal cues. Hence, we introduce both the motion and appearance cues to assist in judging the true location of UAVs.

Concretely, a Global-Local Tracking Framework driven by both Motion and Appearance (GLTF-MA) for infrared anti-UAV is proposed in this paper, which consists of the following four modules. Firstly, a Periodic Global Detection (PGD) module is periodically performed to use a global detector to re-locate UAVs in the whole image, which alleviates the problems of UAV frequent appearance/disappearance and unstable flight paths. Meanwhile, a Multi-stage Local Tracking (MLT) module is routinely implemented to deal with the tiny size of UAVs and background interference, which contains a Priori Stage Switching (PSS) mechanism, Motion-Appearance Matching mechanism (MAM), and a Motion Estimation Punisher (MEP). Especially, the PSS mechanism uses an efficient local tracker [12] to give the initial target box, followed by calculating its final score with prior knowledge to accurately determine whether the local tracker is misleading. When the PSS mechanism thinks that the local tracker is misleading, the MAM mechanism and the MEP are performed to utilize the optical flow and appearance information to locate target. Next, a Target Disappearance Judgement (TDJ) module and a Bounding Box Refinement (BBR) module are performed to give robust target disappearance flags and refine the target box, respectively. The visualization results of GLTF-MA are shown in Fig. 1.

The main contributions of GLTF-MA are summarized as follows.

- A unified tracking framework GLTF-MA is proposed that adaptively switches between local tracking and global detection to deal with UAV frequent appearance/disappearance and unstable flight paths.
- A PGD module containing a PSS mechanism, a MAM mechanism, and a MEP is proposed to solve the tiny size of UAVs and background interference.
- A TDJ module and a BBR module are proposed to give robust target disappearance flags and refine the target box.
- Extensive experimental results show that GLTF-MA performs significantly better than other competing trackers in the 3rd Anti-UAV Challenge.

2. Related Works

2.1. Tracking Paradigm

At present, the architecture of the advanced trackers mainly consists of three parts: a backbone to extract features, a fusion module for aggregating features of the template and the search region, and a prediction head to locate the target. Most trackers regard the modified ResNet [13] as the first choice for backbone, while LightTrack [14] automatically searches the backbone by the one-shot neural architecture search technique. Siamese-based trackers [8–10, 15, 16] use a correlation operation in the fusion module to establish relationships between templates and search regions, while recent transformer-based trackers [12, 17–21] use the attention mechanism in the fusion module to fully utilize the global context information and adaptively focus on useful tracking information. Common candidate prediction heads include the regression and classification-based head [17, 20, 22], the query-based head [23], and the corner-based box estimation head [12, 18].

Although the above tracking pipeline continues to have ingenious modules being proposed, it still faces several major problems in the practical application of anti-UAV. Firstly, as the UAVs continue to move and change the angle, the template feature extracted by the backbone is not in the appearance state that meets the current target. Secondly, because UAVs are usually small targets with fast speed, the search region features contain less target information, which can easily lead to tracking failures. Finally, UAVs are often obscured due to fluttering clouds, and this tracking paradigm cannot determine when UAVs disappear and how to recover them in time when UAVs reappear in view. As a result, we design a global-local tracking frame to switch between the local search region and the entire image for robust tracking.

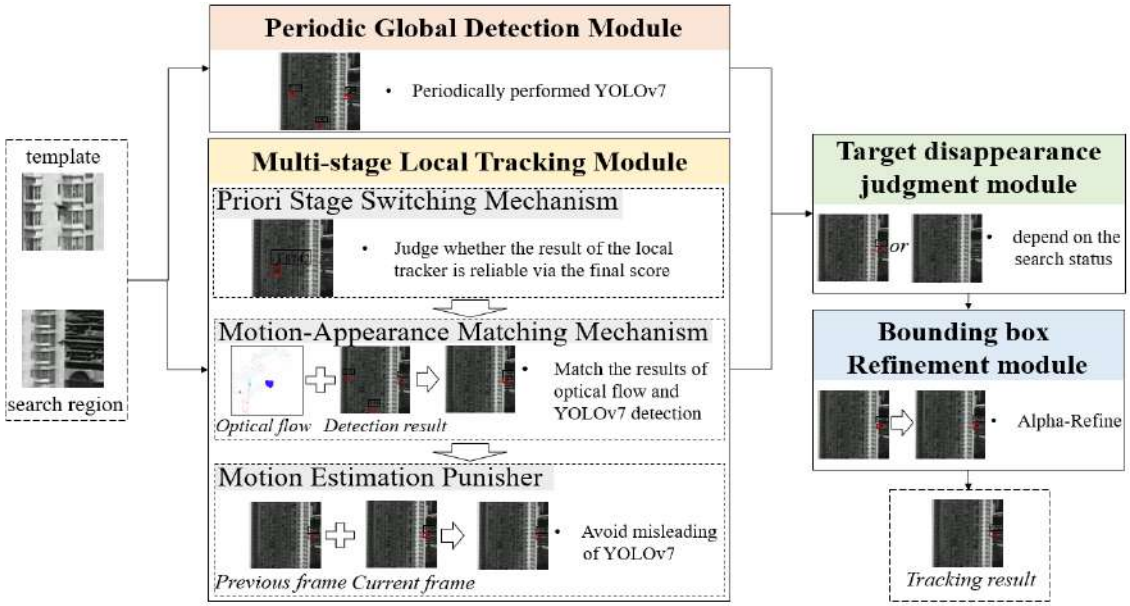


Figure 2. Visualization results of GLTF-MA (ours) and MixFormer on the 3rd Anti-UAV Challenge.

2.2. Tracking by Detection

Many trackers adopt the tracking by detection approach to cope with frequent disappearance/appearance of the target in the view, especially in long-term tracking. TLD [24] is a classical algorithm for solving long-term tracking problems, which uses the optical flow technique and weak classifiers to combine local tracking and global redetection. Based on this, many models [25, 26] that combine different local trackers and different global detectors to solve long-term tracking problems have been proposed. Meanwhile, other models [27, 28] use the mechanism of expanding the search region to locate targets when tracking fails.

Inspired by the tracking by detection approach, on the one hand, we introduce global detection in the PGD module to ensure that the local tracker tracks in the correct search region. On the other hand, we add a MAM mechanism with the global detection capability to the MLT module to deal with the failure of the local tracker.

2.3. Tracking against Background Interference

During the tracking process, the target deformation caused by rotation and scale change will result in the tracker to be interfered with similar-looking objects in the background. To solve this problem, some models [29, 30] have introduced an online template update mechanism to timely capture the target slight appearance changes and avoid interference. STMTrack [31] proposes a space-time memory network to store the historical information of the target and guide the tracker to focus on the area with the most target information. SwinTrack [32] embeds historical target trajectories into motion tokens to improve tracking. Furthermore, some researchers have proposed the fast transformation learning model [33], the distractor-aware module [28], and backward gradients integration [34] to customize the template update method [33–36].

In the context of infrared anti-UAV, the available appearance information is scarce due to the tiny size and blurred appearance, making trackers more susceptible to be interfered. In this paper, we propose the MLT module to alleviate this problem, where the MLT module first uses the PSS mechanism with UAV flight prior knowledge to prevent the local tracker from mis-tracking, then adopts the MAM mechanism to suppress interference by matching the optical flow results and global detection results. Finally, a MEP is used to further ensure that the search region remains within acceptable limits.

3. GLTF-MA

In this section, GLTF-MA is introduced in detail, which consists of four modules: a PGD module, an MLT module, a TDJ module, and a BBR module, whose overall framework is presented in Fig. 2. Firstly, the PGD module is performed periodically to ensure that the target is included in the search region of the MLT module, which deals with the frequent appearance/disappearance of targets and erratic flight paths. The MLT module is executed almost every frame, which performs corresponding search techniques stage by stage based on the current tracking reliability. Then, the TDJ module judges whether the target is out of view according to the tracking status. Finally, when the target is considered

to be in view, the BBR module adjusts the outputs of the PGD and MLT modules to give more accurate target boxes.

3.1. Periodic Global Detection Module

Existing trackers often crop the current search region according to the target location in the previous frame, whose performance strongly depends on the accuracy of the prediction location in the previous frame. However, UAVs have the characteristics of frequent appearance/disappearance in the view and erratic flight paths, which greatly reduces the correlation between the current target position and the last target position, making existing trackers fail. To solve this problem, we adopt the PGD module to re-locate the target within the full image to ensure the search region of the MLT module contains the target. Here, YOLOv7 [37] is adopted as the global detector due to its fast inference speed and high detection accuracy. Compared with other YOLO variants [38–40], YOLOv7 designs trainable bag-of-freebies methods, so that real-time object detection can greatly improve the detection accuracy without increasing the inference cost. In addition, YOLOv7 proposes "extend" and "compound scaling" methods for the real-time object detector that can effectively utilize parameters and computation. As a result, we adopt YOLOv7 as the base detector in the PGD module.

Specifically, we firstly use the UAVs in the training set of the 3rd Anti-UAV Challenge as a new class to fine-tuning the trainable parameters of YOLOv7. Then, the trained YOLOv7 performs global UAV detection on the entire image every I frame. If the confidence score of the best box in the detection result is greater than the threshold δ, the box is selected as the tracking result of the current frame. Otherwise, the MLT module is performed to search for the target position.

3.2. Multi-stage Local Tracking Module

The UAV flight background is complex and the appearance information is scarce, making a single local tracker easy to lose the target. In this paper, we propose the MLT module including a PSS mechanism, a MAM mechanism, and a MEP to solve this problem, which are specifically introduced as follows.

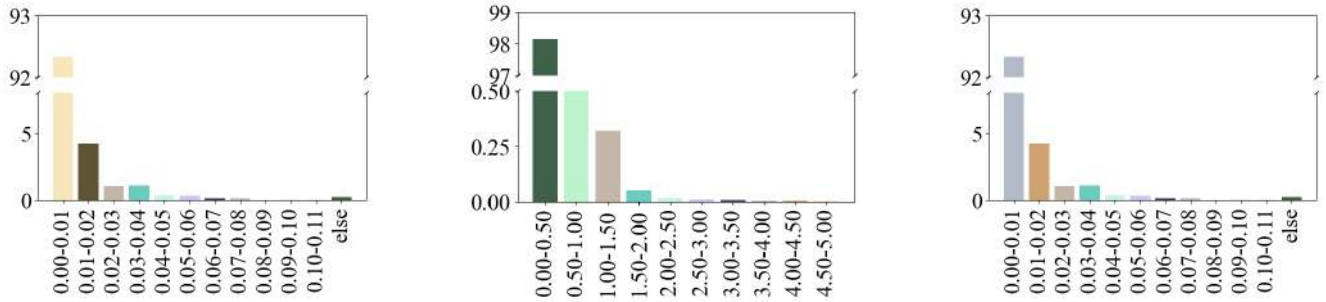
Priori Stage Switching mechanism. Firstly, the template and the search region of frame t are input into the local tracker, i.e., MixFormer [12], to obtain an initial target box S_t^{loc} with its SPM score [12]. S_t^{loc} consists of a four-tuple $[x_t^{loc}, y_t^{loc}, w_t^{loc}, h_t^{loc}]$, where x_t^{loc} and y_t^{loc} represent the center of the box, w_t^{loc} and h_t^{loc} mean the weight and the height of the box. The SPM score is derived from the SPM branch of MixFormer, which represents the similarity between the appearance of the initial template and the region contained in S_t^{loc}. However, as shown in Fig. 5, when the tracker fails to track, the SPM score is still high, which may be because the low resolution of the image with limited UAV appearance information misleads the judgment of the SPM branch.

To deal with this problem, we observe the statistical properties of UAVs in the training set of the 3rd Anti-UAV Challenge. As shown in Fig. 3, compared with the target box in the previous frame, the absolute area, aspect ratio, and center position of the box in the current frame have little change. Hence, we introduce these prior knowledge to calculate the final score F_s of S_t^{loc} to judge whether S_t^{loc} actually tracks the target and whether to switch to the next tracking stage.

$$F_s = SPM - 100 * (4 * P_{area} - P_{move} - \frac{P_{ratio}}{1000}) \quad (1)$$

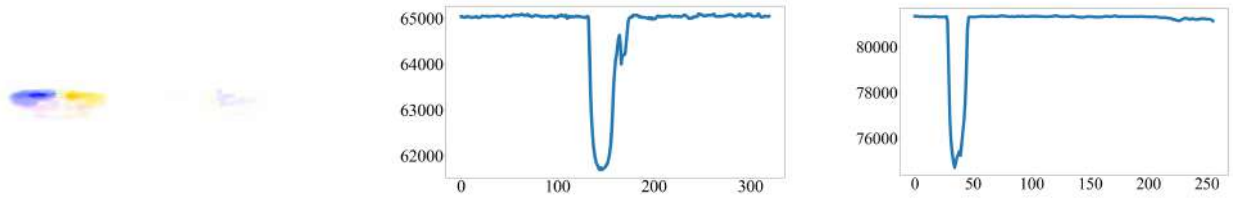
where $P_{area} = |w_t^{loc} * h_t^{loc} - w_{t-1}^{loc} * h_{t-1}^{loc}|$ is the absolute area penalty, $P_{move} = |dis([x_t^{loc}, y_t^{loc}])/dis([0,0],[w,h])|$ is the movement distance penalty (w and h refer to the weight and height of the entire image, $dis(\cdot,\cdot)$ means the Euclidean distance of two points), and $P_{ratio} = |\frac{w_t^{loc}}{h_t^{loc}} - \frac{w_{t-1}^{loc}}{h_{t-1}^{loc}}|$ is the ratio change penalty. If the value of F_s is greater than the threshold α, it is considered that S_t^{loc} is reliable, and S_t^{loc} is directly input the BBR module, marked as *case 1*. Otherwise, the MAM is activated to continue tracking.

Motion-appearance Matching mechanism. Once the MAM mechanism is activated, local tracking only with appearance information is unreliable. To make full use of the temporal and spatial information inherent in the videos, on the one hand, we use the global detector (i.e., YOLOv7) to detect UAVs on the whole image, which may detect multiple target boxes marked as $\{S_t^{glo_1}, S_t^{glo_2}, ..., S_t^{glo_n}\}$. On the other hand, we use frame $t-1$ and frame t to calculate the dense optical flow [41] to find the most prominent motion location l_t^{of}. As shown in Fig. 4, $l_t^{of} = \{x^{of}, y^{of}\}$ is a two-tuple representing the pixel position, where x^{of} and y^{of} are the position indexes where the sum of pixel values along the y-axis and the x-axis are the smallest on the optical flow map, respectively. Next, $\{S_t^{glo_1}, S_t^{glo_2}, ..., S_t^{glo_n}\}$ and l_t^{of} are matched to find the region most likely to be the target. Specially, if $n > 0$, we find the closest box in $\{S_t^{glo_1}, S_t^{glo_2}, ..., S_t^{glo_n}\}$ to l_t^{of} and mark it as $S_t^{glo_{best}}$, where the shortest distance is marked as d_t^{glo}. Otherwise, we calculate the distance between l_t^{of} and S_{t-1} (the target box in frame f_{t-1}) and mark it as d_t^{of}. However, extensive experiments show than locating the target based on the matching result of $S_t^{glo_{best}}$ and l_t^{of} are not absolutely reliable, such as none of the boxes containing the UAV in the results of YOLOv7 as shown in Fig. 6. Hence, it is necessary to introduce a motion estimation punisher to further avoid tracking failures.



(a) Absolute area (b) Aspect ratio (c) Movement distance

Figure 3. The change statistical results of absolute area, moving distance, and aspect ratio between adjacent frames of UAVs, where the horizontal axis represents the change interval of the corresponding item, and the vertical axis represents the proportion (%) of the frames in this interval to the total number of frames.



(a) Optical flow map (b) Add pixel values along the x-axis (c) Add pixel values along the y-axis

Figure 4. Schematic diagram of the optical flow saliency region calculation. The pixel values of the optical flow map (a) are added along the x-axis and the y-axis to obtain (b) and (c), where the indices of the smallest values determine the most prominent motion location.

Motion Estimation Punisher. To avoid the motion-appearance matching mechanism from misleading subsequent tracking, a MEP that utilizes the moderate target movement between adjacent frames is added. Specially, when the number of YOLOv7 detection results is greater than 0 in the MAM mechanism, if $d_t^{glo} < 0.2 * dis([0,0],[w,h])$, $S_t^{glo_{best}}$ is input into the BBR module, marked as *case 2*. Otherwise, S_{t-1} is input into the BBR module, marked as *case 3*. When the number of YOLOv7 detection results is equal to 0, if $d_t^{of} < 0.2 * dis([0,0],[w,h])$, d_t^{of} is input into the BBR module, marked as *case 4*. Otherwise, S_{t-1} is input into the BBR module, marked as *case 5*.

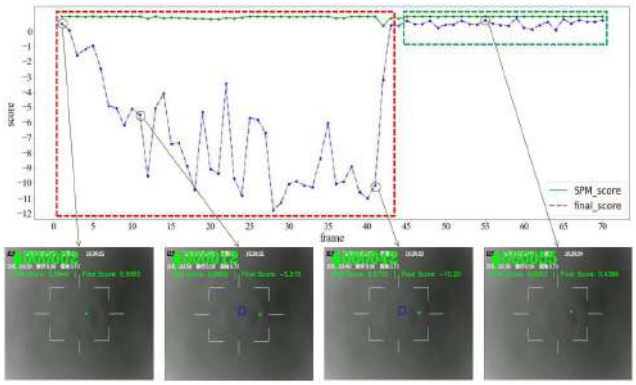


Figure 5. Comparison of the SPM score and the final score in the same video sequence, where the blue box and the green box represent the current frame tracking result and the groundtruth box, respectively. Regardless of whether the tracker tracks the target, the SPM score is always high, while the final score decreases significantly when the tracker fails.

Figure 6. Schematic diagram of the matching between the optical flow map and the detection results of YOLOv7, where $S_t^{glo_2}$ is the closest to l_t^{of}, yet $S_t^{glo_2}$ does not contain the target.

3.3. Target Disappearance Judgement Module

Current trackers are usually able to output a target box with a corresponding confidence score, but do not have the ability to distinguish whether the target is out of view. Extensive experiments show that *cases 1-4* in the PGD and the MLT modules are reasonable, while *case 5* are often biased

Table 1. The whole performance of different trackers on 3rd Anti-UAV valid set. The first-, second- and third-place trackers are labeled with red, blue, and green colors respectively. Best viewed in color.

Trackers	SiamBAN	SiamCAR	SiamFC	SiamMASK	SiamRPN	SiamRPN++	STARK	TransT	OsTrack	MixFormer	Ours
ACC	0.1112	0.2042	0.1545	0.1393	0.1549	0.1709	0.2217	0.2405	0.3045	0.2943	0.4961

from the target. As a result, when more than θ consecutive frames are *case 5*, the target is considered out of view in the current frame.

3.4. Bounding Box Refinement Module

Due to the fuzzy boundary and lack of texture information of infrared UAVs, the results obtained by the tracker suffer from limited accuracy. To furthermore improve the tracking performance, we add the BBR module after the TDJ module to fine tune the target box, where the Alpha-Refine [42] is used as the main component in the BBR module due to its convenience.

The core components of Alpha-Refine include a pixel-wise correlation, a corner prediction head, and an auxiliary mask head. Alpha-Refine is a flexible and accurate refinement plugin, which extracts and maintains detailed spatial information and can significantly improve the quality of the prediction boxes.

3.5. Training and Inference

Training. The training process of GLTF-MA consists of two parts. Firstly, the local tracker, i.e., MixFormer, is fine-tuned by the training set of the 3rd Anti-UAV Challenge, whose training process is the same as [12] except that the epochs of the first stage and the second stage are 50 and 10, respectively. Secondly, the global detector, i.e., YOLOv7, is trained by creating a new class of UAVs, where the training data is derived from UAV images in the training set of the 3rd Anti-UAV Challenge and the training process is the same as [37].

Template Selection and Update. During the inference process of MixFormer, multiple templates and the search region are used together for feature extraction and interaction, but the SPM score used to select the online template is biased. As a result, we use the final score to replace the SPM score in GLTF-MA, i.e., selecting the online template with the highest final score to substitute the previous template within the interval of 200 frames.

Inference. During inference, the PGD module is activated every I frames to perform global detection. If the confidence score of the best box from the PGD module is greater than the threshold δ, the MLT module is skipped and the box is fed into the TDJ module. Otherwise, one initial template, two online templates and a search region are fed into the MLT module to produce the target box with different cases. When the TDJ module thinks that there is a target in the filed of view, the BBR module refines the target box to get a more accurate bounding box. Otherwise, GLTF-MA directly gives that the target is invisible.

4. Experiments

4.1. Experimental Setup

Trackers. In the field of visual object tracking, Siamese-based and Transformer-based trackers have gained increasing attention due to their high tracking accuracy and robustness. In this section, six classic Siamese-based trackers (SiamBAN [16], SiamCAR [15], SiamFC [8], SiamMask [43], SiamRPN [9], SiamRPN++ [10]) and four Transformer-based trackers (STARK [18], TransT [17], OsTrack [11], MixFormer [12]) are introduced. By comparing our proposed method with these ten state-of-the-art trackers, we aim to demonstrate the effectiveness and superiority of our approach in the anti-UAV task. Comprehensive experimental results on anti-UAV benchmarks will be presented to validate our claims.

Evaluation Metrics. To exhaustively analyze the performance of trackers, we evaluate the performance of all frames as follows:

$$acc = \sum_{t=1}^{T} \frac{IoU_t \times \sigma(v_t > 0) + p_t \times (1 - \sigma(v_t > 0))}{T} - 0.2 \times (\sum_{t=1}^{T^*} \frac{p_t \times \sigma(v_t > 0)}{T^*})^{0.3} \tag{2}$$

where for frame t, IoU_t is intersection over union between the predicted box and the ground-truth box. p_t is the predicted visibility flag, which equals 1 when the predicted box is empty and 0 otherwise. v_t is the ground-truth visibility flag of the target, the indicator function $\sigma(v_t > 0)$ equals 1 when $v_t > 0$ and 0 otherwise. The accuracy is averaged over all frames in a sequence, T means total frames, and T^* denotes the number of frames corresponding to the presence of the target in the ground-truth.

Parameters. We choose the MixVit-L version of Mixformer with ConvMAE pre-training as our local tracker. The sizes of the search region and templates are set to 384 × 384 pixels and 192 × 192 pixels, respectively. For the global detection in the PGD module, the re-detection interval I and the box confidence threshold δ are set to 15 and 0.55, which determines whether the detection is reliable. The threshold

α for the final score is set to -3.0, and the result of the local tracking is considered reliable when the final score is greater than α. In the TDJ module, when more than $\theta = 15$ consecutive frames are *case 5*, the target is considered out of view.

Table 2. Illustration of attribute annotation in ANTI-UAV.

Attribute	Description
OV	Out-of-View: the target leaves the view.
OC	Occlusion: the target is partially occluded or heavily occluded.
FM	Fast Motion: the ground-truth's motion between two adjacent frames is larger than 60 pixels.
SV	Scale Variation: the ratio of the bounding boxes of the first frame and the current frame is out of the range [0.66, 1.5].
LI	Low Illumination: the illumination in the target region is low.
TC	Thermal Crossover: the target has a similar temperature with other objects or background surroundings.
LR	Low Resolution: the number of pixels inside the ground-truth bounding box is less than 400 pixels.

4.2. Comparison Studies

We compare GLTF-MA with existing state-of-the-art trackers on the 3rd Anti-UAV Challenge and the ANTI-UAV benchmark [44], and these benchmarks and comparison results are introduced as follows.

The 3rd Anti-UAV Challenge. The validation set of the 3rd Anti-UAV Challenge consists of 50 sequences with various backgrounds such as sky, buildings, and mountains. At the same time, the UAVs may be stationary, moving quickly, low resolution, obscured, and out of sight. As shown in Table 1, GLTF-MA performs best with the ACC of 0.4961. Compared with OsTrack and MixFormer, GLTF-MA leads to 0.1961 and 0.2018 improvement on average tracking accuracy.

The valid set of ANTI-UAV. To further compare different trackers on various attributes, we observe the performance of each tracker on the valid set and the test set of ANTI-UAV with attribute annotations, where the attributes are described in Table 2. In the validation set consisting of 67 video sequences of ANTI-UAV, GLTF-MA outperforms other trackers in multiple attributes, including OV, OC, FM, SV, LI, TC, and LR, whose specific results are shown in Table 3. Particularly, GLTF-MA exhibits a significant performance improvement in LR and FM. Considering the average score of multiple attributes, GLTF-MA achieves the best performance with the ACC of 0.6556.

The test set of ANTI-UAV. As shown in Table 4, in the test set consisting of 91 video sequences of ANTI-UAV, GLTF-MA exhibits the best performance on OV, FM, SV, LI, TC, and LR attributes, which improves the ACC value by 0.2241, 0.1896, 0.1819, 0.2882, 0.1040, and 0.3026 compared to the second-place trackers. Only in the OC attribute, the performance of GLTF-MA is slightly worse than that of MixFormer. It can be concluded that GLTF-MA can achieve effective tracking in various complex scenarios.

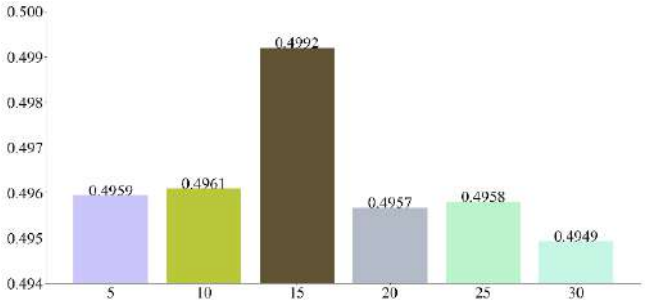


Figure 7. Performance comparison of GLTF-MA with different I.

4.3. Exploration Studies

To verify the effectiveness and give a thorough analysis on GLTF-MA, we perform a series of exploration studies on the valid set of the 3rd Anti-UAV Challenge.

Study on the building modules. To illustrate the effectiveness of our designed tracking framework, we sequentially add the key modules on the base tracker, i.e., MixFormer, to observe their performance. As shown in the first four rows of Table 5, the ACC of the original MixFormer is 0.2943, while after adding PGD, MLT, and BBR modules in sequence, the performance increases to 0.4348, 0.4319, and 0.3673, which shows that these three modules all act an import role on tracking performance. In addition, as shown in the last two rows of Table 5, we compare GLTF-MA without the TDJ module with GLTF-MA, and the performance is improved from 0.4494 to 0.4961, which demonstrates the rationality of the TDJ module to judge whether the target is out-of-view based on the tracking reliability.

Study on the number of intervals for PGD. In GLTF-MA, the execution cycle I of the PGM module determines the frequency of switching between local tracking and global detection. We observe the effect of different I on the tracker, as shown in Fig. 7. The results show that GLTF-MA performs best when $I = 15$, because when the global detection is too frequent, the tracker is easily disturbed by extremely distant backgrounds, while the global detection is too sparse, the local tracker may remain in the wrong tracking state for a long time.

Study on the judgement basis for target disappearance in TDJ. In the TDJ module, it is judged whether the target is out-of-view according to the number of consecutive frames θ of *case 5*. As shown in Fig. 8, when $\theta = 1$, the performance of GLTF-MA is the worst. This may be because the TDJ module misjudges many frames, where these

Table 3. The whole attributed-based performance of different trackers on Anti-UAV valid set. The first-, second- and third-place trackers are labeled with red, blue and green colors respectively.

Trackers	OV	OC	FM	SV	LI	TC	LR	ALL
SiamBAN	0.0059	0.0856	-0.0051	0.0532	0.3617	0.3632	-0.0426	0.1174
SiamCAR	0.0807	0.1533	0.1792	0.0472	0.4851	0.4513	0.0339	0.2043
SiamFC	0.0694	0.0428	0.1346	-0.0043	0.4233	0.4234	-0.0113	0.1539
SiamMASK	0.0117	0.0259	0.0560	-0.0134	0.3467	0.3345	-0.0372	0.1026
SiamRPN	0.0460	0.1245	0.0678	0.0984	0.3976	0.4046	0.0488	0.1696
SiamRPN++	0.0117	0.0259	0.0560	-0.0134	0.3467	0.3345	-0.0372	0.1716
STARK	0.3510	0.3119	0.4055	0.3448	0.5014	0.4999	0.3593	0.3962
TransT	0.1292	0.4938	0.4705	0.3205	0.5795	0.5226	0.3398	0.3989
OsTrack	0.2856	0.5309	0.5633	0.3886	0.6344	0.57460	0.4161	0.4847
MixFormer	0.3667	0.4437	0.5340	0.4146	0.5720	0.5964	0.4841	0.4873
Ours	0.4644	0.6606	0.7178	0.6263	0.7319	0.7053	0.6833	0.6556

Table 4. The whole attributed-based performance of different trackers on Anti-UAV test set. The first-, second- and third-place trackers are labeled with red, blue and green colors respectively.

Trackers	OV	OC	FM	SV	LI	TC	LR	ALL
SiamBAN	-0.0153	0.1820	-0.0035	-0.0068	-0.1446	0.2907	-0.1300	0.0246
SiamCAR	0.1418	0.6065	0.1300	0.1203	-0.0232	0.3082	-0.0595	0.1748
SiamFC	-0.0051	0.5637	0.0392	0.0453	-0.0664	0.1688	-0.1152	0.0900
SiamMASK	0.0466	0.3082	-0.0112	0.0024	-0.1457	0.1982	-0.1278	0.0386
SiamRPN	0.0602	0.3005	0.0602	0.0144	-0.0653	0.2679	-0.1047	0.0761
SiamRPN++	0.0408	0.3206	0.0177	0.0294	-0.1210	0.2988	-0.1066	0.0685
STARK	0.2378	0.5307	0.2797	0.2314	0.1655	0.4179	0.1534	0.2800
TransT	0.1744	0.5921	0.2338	0.1589	0.1155	0.3884	0.0314	0.2420
OsTrack	0.3917	0.6352	0.4320	0.3556	0.4061	0.4843	0.3167	0.4316
MixFormer	0.4012	0.6596	0.4533	0.3966	0.4095	0.5192	0.3111	0.4500
Ours	0.6253	0.6485	0.6429	0.5785	0.6977	0.6232	0.6137	0.6328

Table 5. Ablation for the building modules.

Base Tracker	PGD	MLT	BBR	TDJ	ACC
MixFormer					0.2943
	✓				0.4348 (+0.1405)
		✓			0.4319 (+0.1376)
			✓		0.3673 (+0.0730)
	✓	✓	✓		0.4494 (+0.1551)
Ours	✓	✓	✓	✓	0.4961 (+0.2018)

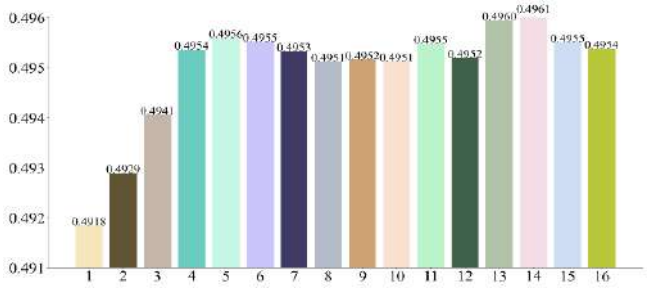


Figure 8. Performance comparison of GLTF-MA with different θ.

frames are only temporarily difficult to track, but the tracker will quickly find the target in subsequent frames. When $3 \leq \theta \leq 16$, the effect of different θ on GLTF-MA is small, so we set $\theta = 15$ in this paper.

5. Conclusion

In this work, we propose a global-local tracking framework driven by motion and appearance for infrared Anti-UAV, i.e., GLTF-MA, including the PGD, the MLT, the TDJ, and the BBR modules. The PGD module periodically re-locates UAVs to ensure that the search region of the local tracker contains the target. Meanwhile, the MLT module performs multi-stage tracking to prevent wrong tracking by the local tracker. Next, the TDJ module is performed to determine whether the target is out of view. At last, the BBR module fine-tunes the target box if the target is present in the field of view. Extensive experimental results on the 3rd Anti-UAV Challenge and the ANTI-UAV benchmark show that GLTF-MA outperforms current state-of-the-art trackers, especially in the case of fast movement and low resolution. In the future, we will delve into improving the performance of GLTF-MA by fusing the complementary information of infrared and visible images.

References

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

[2] Jayol Lee, Min Park, Iksoo Eo, and Bontae Koo. An x-band fmcw radar for detection and tracking of miniaturized uavs. In *CSCI*, 2017. 1

[3] Yue Xiao and Xuejun Zhang. Micro-uav detection and identification based on radio frequency signature. In *ICSAI*, 2019. 1

[4] Bowon Yang, Eric T Matson, Anthony H Smith, J Eric Dietz, and John C Gallagher. Uav detection system with multiple acoustic nodes using machine learning models. In *IRC*, 2019. 1

[5] Weixi Liu, Xiangyong Meng, Weixian Qian, Minjie Wan, and Qian Chen. Infrared small target detection algorithm based on multi-directional derivative and local contrast. In *AOPC*, 2019. 1

[6] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Qixiang Ye, Jianbin Jiao, Zhenjun Han, et al. Anti-uav: a large-scale benchmark for vision-based uav tracking. *IEEE Transactions on Multimedia*, 2021. 1

[7] 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:2108.09909*, 2021. 1

[8] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional Siamese networks for object tracking. In *ECCV*, 2016. 1, 2, 6

[9] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 1, 2, 6

[10] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of Siamese visual tracking with very deep networks. In *CVPR*, 2019. 1, 2, 6

[11] Botao Ye, Hong Chang, Bingpeng Ma, and Shiguang Shan. Joint feature learning and relation modeling for tracking: A one-stream framework. *ECCV*, 2022. 1, 6

[12] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. MixFormer: End-to-End tracking with iterative mixed attention. In *CVPR*, 2022. 1, 2, 4, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[14] Bin Yan, Houwen Peng, Kan Wu, Dong Wang, Jianlong Fu, and Huchuan Lu. LightTrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In *CVPR*, 2021. 2

[15] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In *CVPR*, 2020. 2, 6

[16] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *CVPR*, 2020. 2, 6

[17] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 2, 6

[18] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 2021. 2, 6

[19] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *ECCV*, 2022. 2

[20] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *CVPR*, 2022. 2

[21] Runmin Cong, Qinwei Lin, Chen Zhang, Chongyi Li, Xiaochun Cao, Qingming Huang, and Yao Zhao. Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection. *IEEE Trans. on Image Proc.*, 2022. 2

[22] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Target transformed regression for accurate tracking. *CVPR*, 2021. 2

[23] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[24] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1409–1422, 2011. 3

[25] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming-Hsuan Yang. Long-term correlation tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5388–5396, 2015. 3

[26] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *CVPR*, 2020. 3

[27] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *ICCV*, 2021. 3

[28] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018. 3

[29] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 3

[30] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020. 3

[31] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. STMTrack: Template-free visual tracking with space-time memory networks. In *CVPR*, 2021. 3

[32] Liting Lin, Heng Fan, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *arXiv preprint arXiv:2112.00995*, 2021. 3

[33] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. Learning dynamic siamese network for visual object tracking. In *ICCV*, 2017. 3

[34] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Gradnet: Gradient-guided network for visual object tracking. In *ICCV*, 2019. 3

[35] Tianyu Yang and Antoni B Chan. Learning dynamic memory networks for object tracking. In *ECCV*, 2018. 3

[36] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R Scott. Deformable siamese attention networks for visual object tracking. In *CVPR*, 2020. 3

[37] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 4, 6

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 4

[39] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 4

[40] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 4

[41] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003. 4

[42] Houzhang Fang, Xiaolin Wang, Zikai Liao, Yi Chang, and Luxin Yan. A real-time anti-distractor infrared uav tracker with channel feature refinement module. In *CVPR*, 2021. 6

[43] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 6

[44] N. Jiang, K. Wang, X. Peng, X. Yu, and Z. Han. ANTI-UAV: A large multi-modal benchmark for uav tracking. *arXiv: 2101.08466*, 2021. 7