A Unified Transformer-based Tracker for Anti-UAV Tracking

Qianjin Yu[1,*], Yinchao Ma[1,*], Jianfeng He[2],Dawei Yang[2],Tianzhu Zhang[1,†]
[1,2] University of Science and Technology of China
{sa21010105, imyc}@mail.ustc.edu.cn, {yangdawei,hejf}@mail.ustc.edu.cn,tzzhang@ustc.edu.cn

Abstract

Recently, the need for advanced anti-UAV techniques is increasing due to the rising threat of unauthorized drone intrusion. Object tracking, specifically in thermal infrared (TIR) videos, offers a potential solution to this issue. However, the tracked target often suffers dramatic scale variation, frequent target disappearance, and camera movement which severely influence tracking performance. Therefore, we propose a Unified Transformer-based Tracker, dubbed UTTracker, which contains the following four modules. Firstly, a multi-region local tracking module is designed with temporal cues for tackling target appearance variation and multi-region search for tracking targets in multi proposals. Complementarily, a global detection module is introduced to meet the challenge of target frequent disappearance. Meanwhile, a background correction module is incorporated to align the backgrounds between adjacent frames for alleviating camera movement. Particularly, a dynamic small object detection module for tracking the small target that lacks appearance information. Thanks to the designed modules, our UTTracker can achieve robust UAV tracking in TIR scenarios. Numerous experiments on the 1st and 2nd anti-UAV benchmarks demonstrate the effectiveness of UTTracker. Notably, UTTracker is the foundation of the 2nd-place winning entry in the 3rd Anti-UAV Challenge.

1. Introduction

Unmanned aerial vehicles (UAVs) have become increasingly popular for a variety of real-world applications, such as visual surveillance, biological monitoring, and delivery services [11, 43, 62]. However, there is also a risk of UAV abuse, which could harm society. Therefore, anti-UAV technology has very important practical meanings and urgent research needs. Vision-based approaches are more commonly adopted in anti-UAV tasks compared to other approaches because they offer greater flexibility, better accuracy, and higher efficiency. In addition, the tracking technique in thermal infrared (TIR) mode is a key step for anti-UAV tasks, which is well-suited for all weather conditions and low-light scenarios. In contrast with universal visual tracking, there are some deeper challenges in anti-UAV tasks, including target appearance variation, frequent target disappearance, camera movement, and small-scale target tracking [8, 26, 33, 58].

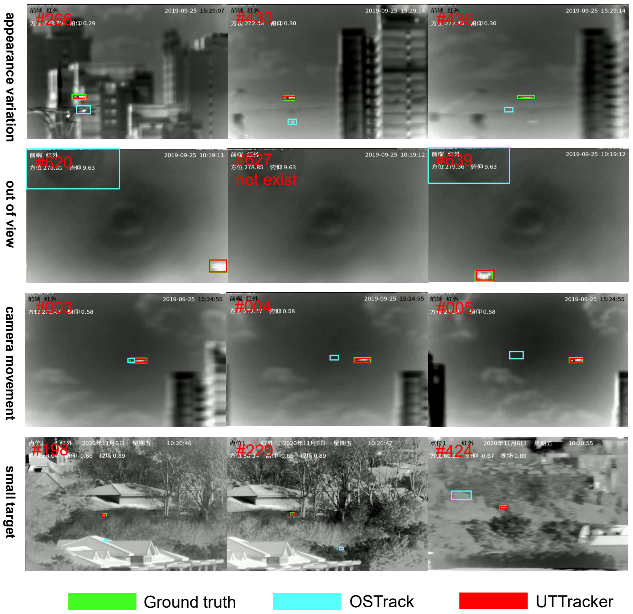

Figure 1. Qualitative comparisons of UTTracker with OStrack baseline on four challenging sequences, with target appearance variation, target out-of-view, camera movement and dynamic small target, respectively. OStrack tends to drift from the target in these scenarios. In contrast, our UTTracker can achieve more robust tracking, owing to the design of MRLT module, GD module, BC module and DSOD module, and thus demonstrates strong robustness in various challenging tracking scenarios.

Over the past few years, object tracking has achieved remarkable progress based on convolutional neural networks [3, 21, 30, 31, 51, 57]. Among them, some works [20, 24, 34, 37, 60] focus on tracking targets in TIR mode to adapt to complex lighting scenarios, providing effective solutions to resolve defects of RGB tracking. However, these CNN-based trackers are difficult to maintain the long-range target relationship between the template and the search region in spatial domains. Recently, thanks to its efficiency and

[*] Equal Contribution
[†] Corresponding Author

ability of global modeling, Transformer has been successfully applied to visual tracking [9, 12, 19, 39, 59]. Typically, MixFormer [12], OSTrack [59], and SimTrack [19] propose plain one-stream tracking frameworks based on Vision Transformers [17, 55] and achieve promising performance. However, these trackers locate the target in a local search region by feature matching, which is difficult to deal with the challenges in the anti-UAV task, as shown in Figure 1.

By studying previous visual tracking methods, We discover the potential of tracking frameworks based on Vision Transformers. To design a special tracker for robust anti-UAV tracking, the following four challenges need to be considered. 1) **Target Appearance Variation**. It is difficult for the existing local trackers to adapt to the variation of target appearance in long-term tracking. Moreover, they are not capable of determining whether the target exists. Thus, we need to improve the local tracker to fit target deformation and improve its ability for target discrimination. 2) **Frequent Target Disappearance**. The common local search strategy is frail to deal with target disappearance and reappearance, because the reappeared target may be beyond the search range. Thus, it is necessary to design a global detection module to locate the target after it disappears. 3) **Camera Movement**. In the anti-UAV task, the camera often moves, which may cause the target to be out of search range. To avoid local tracking loss caused by camera movement, we need to design an effective strategy to align the scenario between adjacent frames. 4) **Small Target Tracking**. Because a small target has little appearance information, tracking methods based on feature matching may be out of work. Therefore, we need to consider small target detection methods that are not based on appearance.

Based on the above discussions, we design a unified transformer-based tracker, termed **UTTracker**, to track UAVs in TIR mode effectively. The UTTracker is composed of four modules, including multi-region local tracking (MRLT), global detection (GD), background correction (BC), and dynamic small object detection (DSOD). In the **Multi-Region Local Tracking** module, we select OSTrack [59] as our Local Tracker (defined as Baseline). Moreover, we introduce a score prediction module (SPM) to determine whether the target exists and design a template update mechanism to adapt to UAV appearance variation. For training a more discriminative SPM, we automatically acquire hard negative samples by tracker instead of random sampling. Further, a multi-region search strategy is integrated into the MRLT module which can detect the target in multiple potential target-existing search regions for robust tracking. In the **Global Detection** module, we contains a global detector to redetect the target after it disappears. If the global detector outputs multi proposals, we can parallelly put all of them into the MRLT module for the final correct location of the target. In the **Background Correction** module, we utilize a dense matching algorithm to align the backgrounds between adjacent frames. By aligning the backgrounds between adjacent frames, our tracker can ensure that camera movements do not cause the target to exceed the search region, resulting in more stable tracking. In the **Dynamic Small Target Detection** module, we propose an improved statistical clustering algorithm that contains morphological operation and dynamic perception range strategy. They are specially designed to detect small-scale UAVs that cannot be tracked using the methods of feature matching. In general, with the collaboration of the four modules, our UTTracker can achieve robust tracking in the TIR model, as shown in Figure 1.

To summarize, the main contributions of this work are: (1) We propose a novel Unified Transformer-based Tracker (UTTracker), which integrates four modules, including Multi-Region Local Tracking, Global Detection, Background Correction, and Dynamic Small Object Detection. It achieves robust TIR UAV tracking. (2) With the combination of MRLT, GD, and BC modules, our tracker can achieve robust tracking in challenging scenarios. (3) To track small targets in complex backgrounds, we design an improved statistical clustering algorithm to capture the small UAVs. (4) We verify the effectiveness of our method by conducting comprehensive experiments on the challenging UAV infrared tracking datasets [28, 63]. Besides, our UTTracker is ranked the 2nd-place in the 3rd Anti-UAV Challenge, which demonstrate our UTTracker achieves the competitive performance in practical scenarios.

2. Related Work

2.1. Single object tracking

Given a target with bounding box annotation in the initial frame of a video, the objective of visual tracking is to localize the target in successive frames. We broadly divide the current popular tracking methods CNN-based and Transformer-based trackers based on network structure that models feature relationships. Among CNN-based trackers [3, 4, 15, 30, 31, 57], SiamFC [3] and SiamRPN [31] utilize convolutional neural networks (CNN) to model the cross-correlation between template and search region features, while DiMP [4] and PrDiMP [15] learn a discriminative target filtering kernel. However, due to the local perceptual limitations of CNN, it is difficult to maintain the long-range target relationship between the template and the search region in spatial and temporal domains. Recently, with the introduction of Transformers in the field of computer vision [6, 17, 40, 55], some transformer-based tracking methods [9, 42, 47, 52, 58] have been proposed. They are proposed as discriminative or Siamese-based trackers, which show better performance than CNN-based trackers because of their long-range feature capture and target discrimination ability. In some early methods [9, 52, 58], CNN was used to extract features separately, where TransT [9] designed a special attention mechanism to realize the target relationship modeling between template and search region features, and TrDiMP [52] learns a global object filter representation across multiple frames based on Transformer. Nowadays, some trackers [7,12,29,56,59] based solely on Transformer architecture have been proposed. OSTrack [59], Mixformer [12], and SimTrack [7] utilize the plain visual Transformers [17, 55] to unify feature learning and relationship modeling, they are better than earlier Transformer-based trackers. However, these trackers use the feature interaction in a local search region to locate the target. Therefore, they lack the ability to track UAVs in TIR mode when occurs target appearance variation, target disappearance, or camera movement. To this end, we need a more well-designed tracker to adapt to the TIR UAV tracking.

2.2. Thermal infrared object tracking

With the rapid development of infrared sensors, thermal infrared (TIR) tracking has received increasing attention due to its ability of thermal spectral images to handle complex scenarios such as darkness, shadows, and illumination changes that are difficult for visual RGB images. Early TIR tracking methods extract hand-crafted features, such as intensity histograms [20], grayscale [22], and gradient histograms [60], from thermal spectral images to perform tracking. Recently, with the development of deep learning, some methods have been widely proposed for more robust TIR tracking. To enhance object feature extraction, pre-trained models on RGB images were utilized to extract deep features. MCFTS [37] constructs an integrated TIR tracking using a VGG feature extraction network with different convolutional features. HSSNet [34] combines multiple layers of convolutional layers and spatially aware networks by combining shallow spatial information with deep semantic features to learn more accurate target discriminative features. SiamSTA [24] introduces a re-detection mechanism that combines local and global search to increase robustness for fast-motion scenes. However, they are difficult to maintain the long-range target relationship between the template and the search region in spatial domains. Meanwhile, we observe several key challenges in the anti-UAV task, such as frequent target disappearance, camera movement, and small targets without significant appearance information. To overcome the above limitation, we propose a novel unified transformer-based tracker to achieve more robust UAV tracking in TIR mode.

2.3. Vision Transformer

Transformer was originally introduced by Vaswani et.al. [50] for machine translation applications, which can establish dependencies of all input tokens and learn representations from a global perspective. The attention mechanism is the vital part of the transformer, which learns the dependencies of all input tokens to aggregate information from the entire input sequence. Over the past year, transformers have shown their great potential in the vision community due to more parallelization and competitive performance. It splits input images into fixed-size patches and then converts them into 1D input tokens. All these tokens are concatenated with a class token and sent into the transformer encoder. The updated class token serves as the global image representation for classification. Nowadays, thanks to its capacity for representation learning, vision transformer has been widely applied in image classification [17,53], object detection [6,55], semantic segmentation [10, 48], action recognition [38, 61], etc. Considering the superiority of vision transformer representational learning, we propose a Unified Transformer-based Tracker (UTTracker) for Anti-UAV Tracking.

3. Method

In this section, we first introduce the overall architecture of our proposed UTTracker. The following four subsections respectively introduce details of the Multi-Region Local Tracking (MRLT) Module, the Global Detection (GD) Module, the Background Correction (BC) Module, and the Dynamic Small Object Detection (DSOD) Module.

3.1. Tracking Architecture

As shown in Figure 2, our UTTracker is composed of four modules to achieve robust UAV tracking in TIR mode, which are the MRLT module, the GD module, the BC module, and the DSOD module. Firstly, the MRLT module selects OSTrack [59] as Local Tracker (defined as Baseline). Based on the Local Tracker, it introduces a template update mechanism and a score prediction module (SPM) for obtaining a dynamic template. Further, the MRLT module adopts a multi-region search strategy that can track the target in multi potential target-existing search regions for discriminative tracking. Secondly, the GD module is incorporated to redetect the target after it disappears, which contains a global detector. Thirdly, we further introduce the BC module to align the backgrounds between adjacent frames. In this way, camera movement will not cause the target to exceed the search region, which facilitates robust tracking. Finally, our UTTracker can automatically enable or disable the DSOD module according to the target scale and the target score, in which we design a Dynamic Small Object Detector for robust small-scale target tracking.

3.2. Multi-Region Local Tracking

In our method, we use a transformer-based tracker as the Local Tracker (defined as Baseline). Then, the template update mechanism and a score prediction module(SPM) are contained to constitute our Local Tracker, as shown in Figure 4. Specifically, we take a triple of images as input for our tracker, including a template image $z \in \mathbb{R}^{3\times H_z\times W_z}$, a

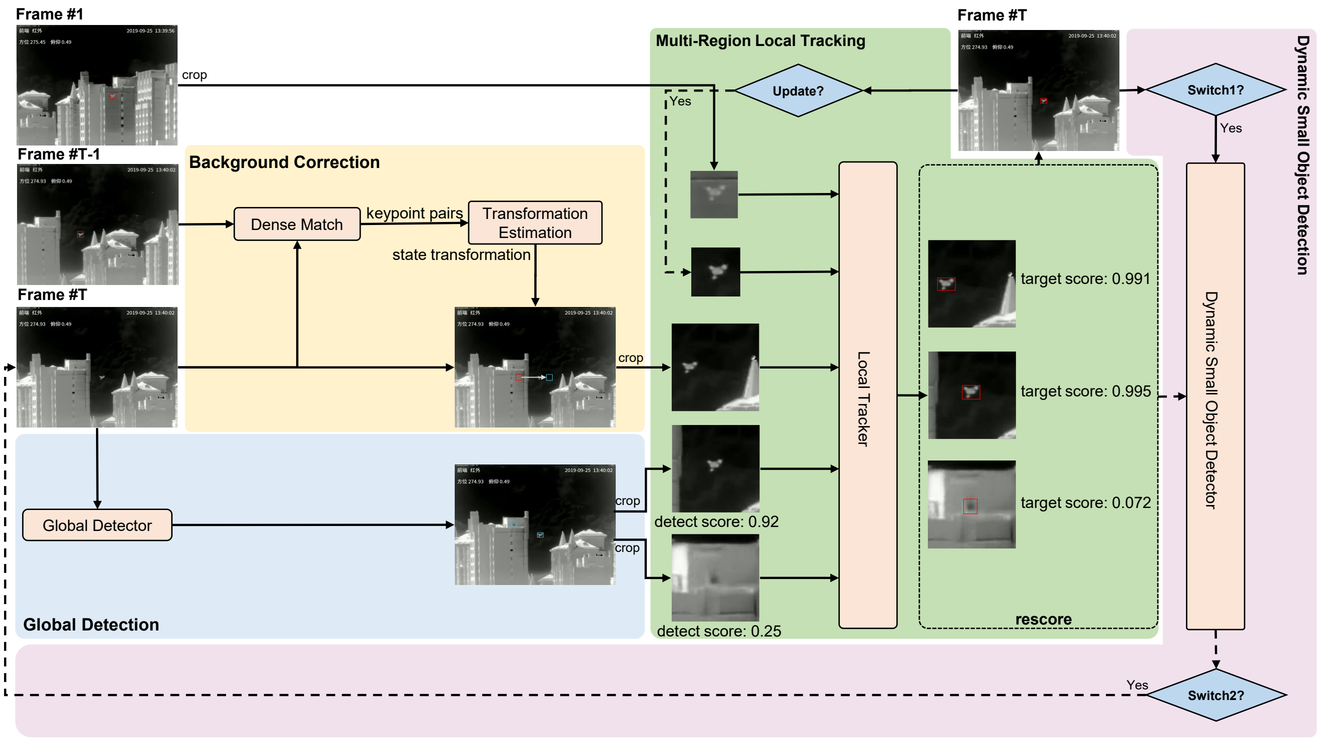

Figure 2. The overall architecture of UTTracker. It contains the MRLT module, the GD module, the BC module, and the DSOD module.

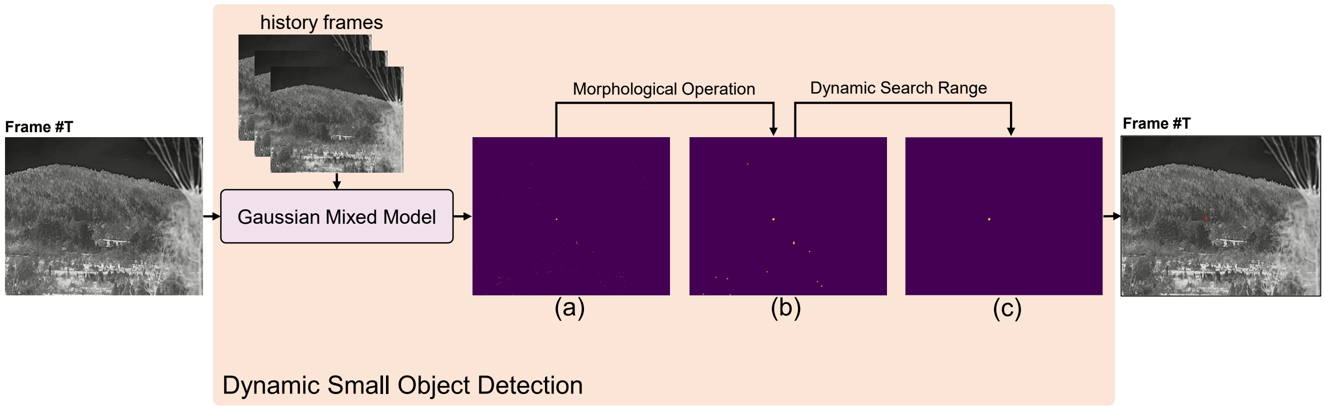

Figure 3. The specific architecture of UTTracker's DSOD module.

dynamic template image $z \in \mathbb{R}^{3\times H_z\times W_z}$, and a search region image $x \in \mathbb{R}^{3\times H_x\times W_x}$. They are split and reshaped into a sequence of flattened 2D patches $\mathbf{z}_p \in \mathbb{R}^{N_z\times(3\cdot p^2)}$, $\mathbf{dz}_p \in \mathbb{R}^{N_z\times(3\cdot p^2)}$ and $\mathbf{x}_p \in \mathbb{R}^{N_x\times(3\cdot p^2)}$, where (p, p) is the resolution of each patch, and $N_z = H_zW_z/p^2$, $N_x = H_xW_x/p^2$ are the patch number of the template and search region respectively. We map flattened patches to C dimension through a linear projection. Then, learnable position embeddings $\mathcal{P}_z \in \mathbb{R}^{N_z\times C}$, $\mathcal{P}_{dz} \in \mathbb{R}^{N_z\times C}$ and $\mathcal{P}_x \in \mathbb{R}^{N_x\times C}$ are added into their corresponding patches to obtain template feature $\mathbf{H}_z^0 \in \mathbb{R}^{N_z\times C}$, dynamic template feature $\mathbf{H}_{dz}^0 \in \mathbb{R}^{N_z\times C}$, and search region feature $\mathbf{H}_x^0 \in \mathbb{R}^{N_z\times C}$. After that, $\mathbf{H}_z^0$, $\mathbf{H}_{dz}^0$ and $\mathbf{H}_x^0$ are concatenated together as input to the vit transformer encoder, which consists of N encoder layers. The encoder layer is composed of multi-head attention and multi-layer perception. Further, we feed the updated search region feature $\mathbf{H}_x^N$ into a box head for target state estimation. Finally, we introduce a score prediction module(SPM) and the template update mechanism to judge the quality of the target and select the high-quality target as the dynamic template.

The SPM we use is the same as MixFormer [12] but has some differences. We automatically acquire negative samples in SPM training. Specifically, we sample a search re-

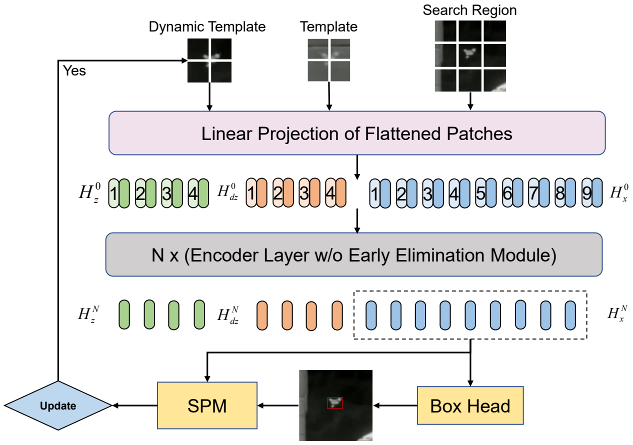

Figure 4. Structure of the Local Tracker with a dynamic template. It contains the score prediction module (SPM) and the template update mechanisms.

gion with the drone (positive region) and a search region without the drone (negative region). We treat the ground truth drone in the positive region as a positive sample and treat the detected target in the negative region as a negative sample. Such negative samples are easier to confuse the SPM, which can help SPM learn better discriminability.

During the tracking process, we design a multi-region search strategy so that targets in multiple cropped search areas can be tracked in parallel. Meanwhile, the SPM outputs target scores in multiple search regions. If the estimated optimal score of SPM is higher than a given target score threshold (e.g. 0.8) and the tracker has reached the update interval (e.g. 20), we resample a template according to the target bounding box corresponding to the optimal score and replace the dynamic template with it. Thanks to the SPM module, template update mechanism, and multi-region search strategy, the MRLT module can greatly adapt to the target scale and appearance change in the process of UAV tracking in TIR mode, which achieves robust tracking.

3.3. Global Detection

Although our MRLT module has achieved stable UAV tracking, when the target is temporarily lost, the local tracker may collapse completely. In order to redetect the target after it disappears, the GD module introduces YOLOv5 [2] as our Global Detector for its convenience and effectiveness. We refer readers to learn more details of our Global Detector's structure in YOLOv5 [2].

In the process of tracking, we first detect the target in each frame. Then the Global Detector outputs all the proposals which might contain the target UAV. We further remove the untrusted proposals whose confidence score is lower than a given global-detect threshold. In the last step, as shown in Figure 2, the multi-region search strategy is employed to rescore the detected proposals that might contain the target. In particular, we crop search regions based on the detected drone proposals and concatenate these regions in a batch. The multi-region local tracking module detects targets in these search regions parallelly and outputs the corresponding target score for selecting the optimal result. If the target scores of the local tracker proposals are lower than 0.5, we judge the drone disappears.

3.4. Background Correction

Equipped with the GD module, our UTTracker fulfills the more robust and precise tracking. However, there are still some common occurrences in drone tracking scenarios, such as camera movement.

During tracking, we design a BC module to align the backgrounds between adjacent frames to avoid local tracking loss caused by camera movement. As shown in Figure 2, the dense matching algorithm, LoFTR [49], is first applied to find key point pairs. More details of the dense matching algorithm can refer to LoFTR [49]. Then, these key points are utilized to regress the optimal single mapping transformation matrix H through the RANSAC [16] algorithm. We align the previous target state $b_{pre} = (x_{pre}, y_{pre}, w_{pre}, h_{pre})$ to the current frame through H. Formally,

$$\begin{bmatrix} x_1 & y_1 & - \\ x_2 & y_2 & - \end{bmatrix} = \begin{bmatrix} x_{pre} & y_{pre} & 1 \\ x_{pre}+w_{pre} & y_{pre}+h_{pre} & 1 \end{bmatrix} \times \mathrm{H}^{\mathrm{T}} \tag{1}$$

$$\mathrm{H} = \begin{bmatrix} h_1 & h_1 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & 1 \end{bmatrix} \tag{2}$$

Then, the aligned previous target state $b_{align} = (x_1, y_1, x_2 - x_1, y_2 - y_1)$ is used to crop the search region for the next frame tracking. In this way, camera movement will not cause the target to exceed the search region, so our BC module facilitates more robust tracking.

3.5. Dynamic Small Object Detection

As aforementioned, with the combination of the MRLT module, the GD module, and the BC module, our UTTracker is powerful to locate targets with abundant appearance information. However, it is difficult to detect small-scale targets through feature matching. The global detector and local tracker tend to output a low target score and drifting box for small targets. To this end, we propose a DSOD module, which employs Gaussian Mixture Model [45] (GMM) to capture the small moving target, as shown in Figure 3. We switch to enable the DSOD module for the next frame if the current target size is smaller than 6×6 and the current target score is lower than 0.5, which is the condition of Switch1, just as you see in Figure 2.

The Gaussian Mixture Model maintains K Gaussian components $\eta\left(X_t, \mu_{i,t}, \Sigma_{i,t}\right)$ for each pixel. Here, $\eta\left(\cdot,\cdot,\cdot\right)$means the Gaussian probability density function, X_t denotes the intensity value of the pixel (x, y), $\mu_{i,t}$ and $\Sigma_{i,t}$ are the mean and variance matrix of component i. Following SiamSTA [24], the Gaussian components can be updated by the input image if pixel value X_t is within 2.5 times the standard deviations of a component. If X_t does not match any of the K components, we determine the pixel is a part of the moving target, as shown in Figure 3(a).

Beyond drones, some scenarios contain dynamic backgrounds, which may confuse drone detection. Thus, we first utilize open morphological operation [46] to suppress the complex background noise. Then, we employ close morphological operation to enhance target perception, which can prevent the target from being divided into many parts mistakenly, as shown in Figure 3(b).

Further, based on the previous target location, we set a perception range of 100 pixels and mask out the candidates beyond the perception range, as shown in Figure 3(c). After that, the candidate with the most similar shape to the target in previous frames is chosen as the tracking drone. Further, to redetect the drone after it has been obscured or disappeared, we design a strategy of dynamic perception range. In particular, the perception range R is enlarged according to the number of the drone disappearing frames N_d, which can be formulated as $R = 100 \cdot e^{0.1 \cdot N_d}$. In this way, we can search for the drone in a larger range after it disappears to avoid the drone exceeding the perception range. Notably, the GMM module is based on pixel-level statistics. If the camera is moving, the GMM will lose efficacy. Therefore, we judge the camera motion through H. If the displacement parameter in H is larger than 1, we judge the camera is moving and initialize GMM with the current frame.

Meanwhile, the local tracker and global detector also give proposals. If the target scores of the proposals are larger than 0.5, we disable the DSOD module, which is the condition of Switch2, as shown in Figure 2.

4. Experiments

4.1. Implementation Details

Our UTTracker are implemented using Python 3.6 and Pytorch 1.7.1. The experiments are conducted on a server with four 24GB NVIDIA RTX 3090 GPUs.

MRLT module details. Our Local Tracker is similar to OSTrack [59] but without the Early Candidate Elimination strategy. Besides, we add the score prediction module (SPM) to our tracker. We crop the template and search region by 2^2 and 4^2 times the target bounding box area respectively. Our tracker takes 256×256 search region image and 128×128 template image as input. Same as OSTrack [59], we first train our model on the training splits of LaSOT [18], GOT-10K [25], COCO2017 [36], and TrackingNet [44], which is a general setting for model training in visual tracking. For fairness, we exclude all UAV classes in these datasets. Common data augmentations, such as translation, horizontal flip, and brightness jittering, are applied in training. The minimal training data unit for our tracker consists of two templates and one search region image, where one template serves as the base template to provide the initial target feature and another template serves as the dynamic template to provide the temporal target feature. Notably, we remove the same sequences in the 2nd Anti-UAV test-dev dataset to rebuild our train dataset based on the 3rd Anti-UAV dataset. Then we fine-tune our tracker on the train dataset which contains 142 high-quality IR video sequences. More tracker settings are discussed in Section 4.4.

Table 1. Comparison with state-of-the-art trackers on 1st and 2nd Anti-UAV test-dev datasets. We report the AUC scores, mean precision, and normal mean precision for each tracker. The best three results are shown in **red**, **blue** and **green** fonts.

Method	1st TestDev [28]			2nd TestDev [63]		
	AUC	P	P_{Norm}	AUC	P	P_{Norm}
UTTracker	**77.9**	**98.0**	**97.7**	**72.4**	**93.4**	**91.7**
Baseline	**73.2**	92.7	**92.1**	**64.2**	83.4	**81.4**
SiamSTA [24]	**72.6**	**96.9**	-	**65.5**	**88.8**	-
SiamRCNN [51]	71.8	**94.8**	-	63.3	**83.7**	-
Globaltrack [26]	70.7	93.2	-	61.1	81.4	-
PrDiMP [15]	60.4	87.2	-	53.6	77.6	-
DiMP50 [4]	57.2	80.1	-	51.6	72.9	-
ATOM [14]	55.8	78.2	-	49.5	70.5	-
KYS [5]	55.4	77.4	-	50.4	71.3	-
STRCF [32]	46.1	62.4	-	40.3	57.5	-
ECO [13]	46.1	62.7	-	41.0	58.6	-
SiamRPN [31]	46.1	62.4	-	41.6	56.8	-
SiamRPN++ [30]	43.5	59.4	-	40.4	56.1	-
AutoTrack [35]	42.8	58.0	-	37.6	53.3	-
ARCF [27]	39.4	54.9	-	32.5	50.0	-
CSRDCF [41]	39.3	55.2	-	33.4	48.5	-
KCF [23]	33.9	47.4	-	26.3	40.2	-

GD module details. We choose YOLOV5x [2] as the initial global detector based on the performance comparison of the different variants. Then, we filter out drone objects smaller than 5×5 in the dataset to avoid noise. Finally, our global detector is also fine-tuned on the train dataset. By the way, we store all detection results in caches to facilitate the high efficiency of tracking in the process of inference.

BC module details. We try different models and eventually choose LoFTR [49] model as our dense matching algorithm. In the same way that the global target detector is used, we store H-matrix parameters between continuous frames for efficient drone tracking.

DSOD module details. We use Gaussian Mixture Model [45] with some ingenious designs such as morphological operation [46] as our DSOD module. They are all implemented by OpenCV.

Table 2. Comparison with state-of-the-art trackers on 1st and 2nd Anti-UAV test-dev datasets in terms of average overlap accuracy. The best three results are shown in **red**, **blue** and **green** fonts.

	SiamRPN++ [30]	SiamRPN [31]	ATOM [14]	KYS [5]	DiMP50 [4]	PrDiMP [15]	Globaltrack [26]	SiamR-CNN [51]	SiamSTA [24]	Baseline	UTTtracker
1st TestDev [28]	44.25	46.82	56.72	56.23	58.14	62.45	72.04	72.95	**74.46**	**74.49**	**80.38**
2nd TestDev [63]	41.02	42.23	50.32	51.22	52.41	55.52	62.15	64.29	**67.30**	**65.26**	**75.13**

Table 3. Top 3 tracker results on the 3rd Anti-UAV test challenge.

Trackers	Score(%)
Anonymous Tracker(1st)	69.7
Our UTTracker(2nd)	68.8
Anonymous Tracker(3rd)	68.0

Table 4. Ablation studies on components of UTTracker.

Trackers	MRLT	GD	BC	DSOD	Score(%)
Baseline	-	-	-	-	65.26
-	√	-	-	-	67.28
-	√	√	-	-	72.49
-	√	√	√	-	73.33
UTTracker	√	√	√	√	**75.13**

4.2. Experimental Setup

Datasets. SiamSTA [24] used the 1st and 2nd Anti-UAV test-dev datasets to evaluate their approach. For fairness, we primarily use the same datasets to evaluate our model. Thanks to the 3rd Anti-UAV challenge, we can further experiment on the 3rd Anti-UAV test-challenge datasets. The 1st Anti-UAV test-dev [28] is a high-quality dataset that contains 100 IR videos and 100 RGB videos.It covers multi-scale UAVs with complex scenarios such as clouds, buildings, etc. The 2nd Anti-UAV test-dev [63] incorporates more complex challenges like tiny objects, and camera movement, which contains 140 IR videos. Compared to the previous challenge, the 3rd Anti-UAV test-challenge dataset adds more challenging video sequences with dynamic backgrounds, and complex movements, such that it covers a greater variety of scenarios with multi-scale UAVs.

Evaluation metrics. First, we use three metrics including area under the curve (AUC) of the success plot, Precision(P), and Normal Precision(P$_{Norm}$) to evaluate, which are widely used in large-scale tracking benchmarks such as LaSOT [18], TrackingNet [44], TNL2K [54], etc. Then, according to the evaluation metric given in the Anti-UAV benchmark [63], we further use average overlap accuracy to assess. It calculates the mean IOU of all videos. For more strict evaluation, when the target exists but isn't tracked, the 3rd Anti-UAV challenge adds a penalty item to the original average overlap accuracy. More details can be found in the 3rd Challenge Page [1].

4.3. Quantitative Evaluation

We evaluate the tracker and compare it with some of the currently best-performing deep trackers on the 1st and 2nd Anti-UAV test-dev datasets, including SiamSTA [24], SiamRCNN [51], Globaltrack [26], PrDiMP [15], DiMP50 [4], ATOM [14], KYS [5], STRCF [32], ECO [13], SiamRPN [31], SiamRPN++ [30], AutoTrack [35], ARCF [27], CSRDCF [41], KCF [23].

The results of the AUC, P, and P$_{Norm}$ which compare the trackers mentioned above on 1st and 2nd AntiUAV test-dev datasets are shown in Table 1. Our UTTracker exceeds other UAV trackers a lot on every test-dev dataset. Specifically, UTTracker obtains the best AUC score of 77.9%, outperforming SiamSTA by **5.3**% on the 1st Anti-UAV test-dev dataset. Meanwhile, UTTracker achieves the top AUC score of 72.4%, surpassing SiamSTA by **6.9**% on the 2nd Anti-UAV test-dev dataset. Our UTTracker can not only perform well during tracking general UAV targets but also deal well with the challenges such as frequent target disappearance, small-scale targets, and camera movement in the 2nd Anti-UAV test-dev dataset.

In addition, Table 2 displays the overall performance of the top 11 trackers in terms of the average overlap accuracy metric defined in Anti-UAV [63]. Our UTTracker gets the top score of 80.38% on the 1st dataset and 75.13% on the 2nd dataset. Compared with SiamSTA [24], UTTracker respectively gets a gain of **5.89**% and **7.83**% on the two datasets. Moreover, in contrast with our baseline model, our UTTracker gets a high gain of **9.87**% on the 2nd dataset. Our extra modules designed for the baseline tracker are effective to solve the complex challenges in the 2nd dataset.

In the end, we report the results on the 3rd Anti-UAV test-challenge datasets. As shown in Table 3, our UTTracker gets second place on this dataset with only a little difference(0.9%) with the champion scheme.

4.4. Ablation Study

In this subsection, we take the ablation study on the 2nd Anti-UAV test-dev [63] to further analyze the effectiveness of design in UTTracker. We adopt average tracking accuracy defined in Anti-UAV as the evaluation criteria according to SiamSTA [24].

Effectiveness of MRLT module. As shown in Table 4, the MRLT module can get a gain of **2.02**% compared to the baseline model. This can be attributed to the template update mechanism and the multi-region search strategy. In addition, Table 5 discovers that the best template update interval is **15**. Meanwhile, that the optimal target-score threshold is **0.5**, as shown in Table 6.

Effectiveness of the GD module. Results in Table 4

Table 5. Performance of different update intervals of the dynamic template.

UpdateVal	Score(%)
5	74.56
10	74.81
15	**75.13**
20	74.78
25	74.75
30	74.63

Table 6. Performance of different target-score thresholds.

Target-Score Thres	Score(%)
0.1	75.02
0.3	75.08
0.5	**75.13**
0.7	75.04
0.9	74.76

Table 7. Performance of different global detectors.

global detector	Score(%)
YOLOv5S	72.35
YOLOv5M	73.28
YOLOV5L	74.07
YOLOv5X	**75.13**

Table 8. Performance of different global-detection thresholds.

Global-Detection Thres	Score(%)
0.2	74.91
0.4	75.03
0.6	**75.13**
0.8	74.48

show the effects of GD module leads to **5.21%** improvement. It's the best among all four components. This is due to the precise switch between global detection and local tracking when the out-of-view target is present again. As shown in Table 7, we try different global detectors in the GD module, when using the YOLOv5X detector, our UTTracker obtains the best performance. In addition, global detection threshold decides whether we use the detector during tracking or not. Table 8 shows that the optimal global detection threshold is **0.6**.

Effectiveness of the BC module. As you can see in Table 4, through adding the BC module, the tracking result obtains a gain of **0.84%**, which can be owed to the effective solution to the camera movement challenge during tracking.

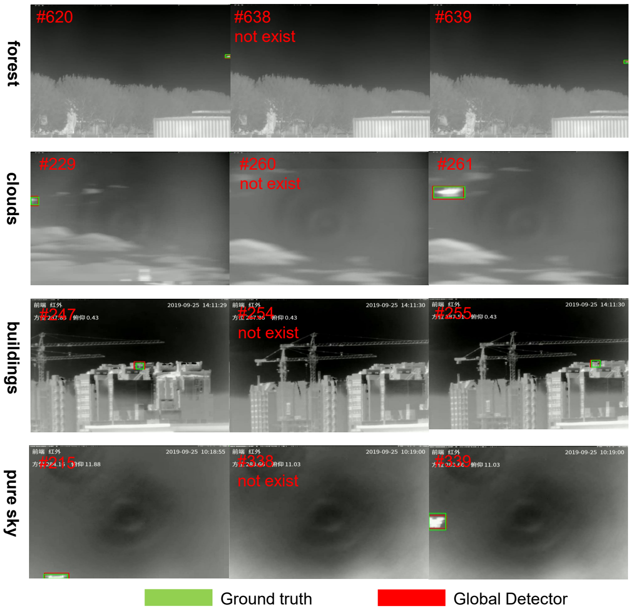

Figure 5. The detection results of our Global Detector when the target disappears and reappears in different scenarios. It can be solved well through our Global Detector.

Effectiveness of the DSOD module. We finally incorporate the DSOD module into our UTTracker. This module is used to solve the challenge of tracking small objects in static backgrounds, which the global detector can't deal well with. As shown in Table 4, the DSOD module brings **1.8%** lift to our UTTracker. This demonstrates that it has the precise perception ability of small-scale targets.

Visualization. As shown in Figure 5, our GD module deals well with the challenge of frequent target disappearance without any sophisticated design. Besides, the more qualitative results of our UTTracker are presented in the Supplementary Materials.

5. Conclusion

In this paper, we propose a robust tracker named UTTracker for Anti-UAV tracking. We first design a multi-region local tracking module concentrating on providing the temporal cues for target appearance variation. Then, we incorporate a global detection module into our UTTracker to redetect the out-of-view target. Meanwhile, we introduce a background correction module to mitigate the effects of camera movement by aligning the backgrounds between adjacent frames. Finally, we design a dynamic small object detection module for stable tracking of the small target without abundant semantic information. With the cooperation of the four modules, our UTTracker achieves the best performance on the 1st & 2nd Anti-UAV benchmarks. Meanwhile, our UTTracker is ranked the 2nd-place in the 3rd Anti-UAV Challenge. However, there are still some shortcomings in our work, such as tracking small targets in moving background, so we will improve our algorithm for more robust UAV tracking in the future.

References

[1] 3rdanti-uav. https://anti-uav.github.io/Evaluate/. 7

[2] Yolov5. https://github.com/ultralytics/yolov5. 5, 6

[3] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip H S Torr. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision Workshops*, 2016. 1, 2

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 6, 7

[5] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *European Conference on Computer Vision*, pages 205–221. Springer, 2020. 6, 7

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 3

[7] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *Proceedings of the European Conference on Computer Vision*, 2022. 3

[8] CL Philip Chen, Hong Li, Yantao Wei, Tian Xia, and Yuan Yan Tang. A local contrast method for small infrared target detection. *IEEE transactions on geoscience and remote sensing*, 52(1):574–581, 2013. 1

[9] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 3

[11] Oliver M Cliff, Debra L Saunders, and Robert Fitch. Robotic ecology: Tracking small dynamic animals with an autonomous aerial vehicle. *Science Robotics*, 3(23):eaat8409, 2018. 1

[12] Yutao Cui, Jiang Cheng, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 4

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient convolution operators for tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6, 7

[14] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 6, 7

[15] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 6, 7

[16] Konstantinos G Derpanis. Overview of the ransac algorithm. *Image Rochester NY*, 4(1):2–3, 2010. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. 2, 3

[18] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A high-quality benchmark for large-scale single object tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 6, 7

[19] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 13774–13783, 2021. 2

[20] Shu Juan Gao and Seong Tae Jhang. Infrared target tracking using multi-feature joint sparse representation. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, pages 40–45, 2016. 1, 3

[21] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[22] Yu-Jie He, Min Li, JinLi Zhang, and Jun-Ping Yao. Infrared target tracking via weighted correlation filter. *Infrared Physics & Technology*, 73:103–114, 2015. 3

[23] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014. 6, 7

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1212, 2021. 1, 3, 6, 7

[25] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 6

[26] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11037–11044, 2020. 1, 6, 7

[27] Ziyuan Huang, Changhong Fu, Yiming Li, Fuling Lin, and Peng Lu. Learning aberrance repressed correlation filters for real-time uav tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2891–2900, 2019. 6, 7

[28] 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. 2, 6, 7

[29] Jin-Peng Lan, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Xu Bao, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. Procontext: Exploring progressive context transformer for tracking. *arXiv preprint arXiv:2210.15511*, 2022. 3

[30] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 6, 7

[31] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 6, 7

[32] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4904–4913, 2018. 6, 7

[33] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 1

[34] 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. 1, 3

[35] Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, and Geng Lu. Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11923–11932, 2020. 6, 7

[36] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014. 6

[37] 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. 1, 3

[38] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022. 3

[39] Yuanpei Liu, Xingping Dong, Wenguan Wang, and Jianbing Shen. Teacher-students knowledge distillation for siamese trackers. *arXiv preprint arXiv:1907.10586*, 2019. 2

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[41] Alan Lukezic, Tomas Vojir, Luka ˇCehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6309–6318, 2017. 6, 7

[42] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8731–8740, 2022. 2

[43] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 445–461. Springer, 2016. 1

[44] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision*, 2018. 6, 7

[45] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009. 5, 6

[46] Khairul Anuar Mat Said, Asral Bahari Jambek, and Nasri Sulaiman. A study of image processing using morphological opening and closing processes. *International Journal of Control Theory and Applications*, 9(31):15–21, 2016. 6

[47] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8791–8800, 2022. 2

[48] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7262–7272, 2021. 3

[49] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 5, 6

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances of Neural Information Processing Systems*, 2017. 3

[51] Paul Voigtlaender, Jonathon Luiten, Philip H. S. Torr, and Bastian Leibe. Siam R-CNN: Visual tracking by re-detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 6, 7

[52] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1571–1580, 2021. 2, 3

[53] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 568–578, 2021. 3

[54] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021. 7

[55] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 22–31, 2021. 2, 3

[56] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8760, 2022. 3

[57] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 1, 2

[58] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10448–10457, 2021. 1, 2

[59] Botao Ye, Hong Chang, Bingpeng Ma, and Shiguang Shan. Joint feature learning and relation modeling for tracking: A one-stream framework. *Proceedings of the European Conference on Computer Vision*, 2022. 2, 3, 6

[60] Xianguo Yu, Qifeng Yu, Yang Shang, and Hongliang Zhang. Dense structural learning for infrared object tracking at 200+ frames per second. *Pattern Recognition Letters*, 100:152–159, 2017. 1, 3

[61] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. *Proceedings of the European Conference on Computer Vision*, 2022. 3

[62] Suofei Zhang, Xu Cheng, Haiyan Guo, Lin Zhou, and Zhenyang Wu. Tracking deformable parts via dynamic conditional random fields. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 476–480. IEEE, 2014. 1

[63] 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. 2, 6, 7