International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01937-0

BioDrone: A Bionic Drone-Based Single Object Tracking Benchmark for Robust Vision

Xin Zhao[1,2] · Shiyu Hu[2] · Yipei Wang[3] · Jing Zhang[2] · Yimin Hu[4] · Rongshuai Liu[4] · Haibin Ling[5] · Yin Li[6] · Renshu Li[4] · Kun Liu[4] · Jiadong Li[4]

Received: 16 December 2022 / Accepted: 17 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Single object tracking (SOT) is a fundamental problem in computer vision, with a wide range of applications, including autonomous driving, augmented reality, and robot navigation. The robustness of SOT faces two main challenges: *tiny target* and *fast motion*. These challenges are especially manifested in videos captured by unmanned aerial vehicles (UAV), where the target is usually far away from the camera and often with significant motion relative to the camera. To evaluate the robustness of SOT methods, we propose **BioDrone**—the first **bio**nic **drone**-based visual benchmark for SOT. Unlike existing UAV datasets, BioDrone features videos captured from a flapping-wing UAV system with a major camera shake due to its aerodynamics. BioDrone hence highlights the tracking of tiny targets with drastic changes between consecutive frames, providing a new robust vision benchmark for SOT. To date, BioDrone offers the largest UAV-based SOT benchmark with high-quality fine-grained manual annotations and automatically generates frame-level labels, designed for robust vision analyses. Leveraging our proposed BioDrone, we conduct a systematic evaluation of existing SOT methods, comparing the performance of 20 representative models and studying novel means of optimizing a SOTA method (KeepTrack Mayer et al. in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 13444–13454, 2021) for robust SOT. Our evaluation leads to new baselines and insights for robust SOT. Moving forward, we hope that BioDrone will not only serve as a high-quality benchmark for robust SOT, but also invite future research into robust computer vision. The database, toolkits, evaluation server, and baseline results are available at http://biodrone.aitestunion.com.

Keywords Robust vision · Visual tracking · Flapping-wing aerial vehicle · High-quality benchmark · Tracking evaluation system

Communicated by Oliver Zendel.

✉ Xin Zhao
xzhaopersonal@foxmail.com

✉ Shiyu Hu
hushiyu2019@ia.ac.cn

Yipei Wang
220213711@seu.edu.cn

Jing Zhang
jing_zhang@ia.ac.cn

Yimin Hu
ymhu2015@sinano.ac.cn

Rongshuai Liu
rsliu2020@sinano.ac.cn

Haibin Ling
hling@cs.stonybrook.edu

Yin Li
yin.li@wisc.edu

Renshu Li
lirenshu2021@gusulab.ac.cn

Kun Liu
liukun2021@gusulab.ac.cn

Jiadong Li
jdli2009@sinano.ac.cn

1 School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

2 Institute of Automation, Chinese Academy of Sciences, Beijing, China

3 School of Instrument Science and Engineering, Southeast University, Nanjing, China

4 Suzhou Institute of Nano-tech and Nano-bionics, Chinese Academy of Sciences, Suzhou, China

5 Department of Computer Science, Stony Brook University, Stony Brook, USA

6 University of Wisconsin-Madison, Madison, USA

Published online: 02 December 2023

1 Introduction

Single object tracking (SOT) (Huang et al. 2021; Hu et al. 2023), an essential computer vision task that aims to locate a user-specified moving target, has attracted numerous researchers to propose effective tracking algorithms (Henriques et al. 2014; Bertinetto et al. 2016; Li et al. 2018; Mayer et al. 2021; Cui et al. 2022). Although existing methods have been widely used in application scenarios like self-driving (Kong and Fu 2022; Dendorfer et al. 2021), augmented reality (Abu Alhaija et al. 2018; Gauglitz et al. 2011) and robot navigation (Dupeyroux et al. 2019; Ramakrishnan et al. 2021), key challenges like *tiny target* and *fast motion* can still affect the robustness of algorithms. SOT is commonly formulated as a sequential decision process (i.e., tracking the current frame should rely on previous frames' tracking results), and corresponding tracking algorithms highly depend on the target's appearance and motion information during execution. However, the tiny target means that the available appearance information is limited, while fast motion increases the difficulty in modeling motion information, and even the relative movement of the target and camera can disrupt motion continuity. Therefore, building a high-quality environment for researching the aforementioned challenging factors can contribute to enhancing the robustness of trackers.

Regrettably, the majority of SOT datasets are designed for generic scenarios, with a primary focus on addressing generalization issues. Thus, they always encompass a wide range of target categories and scene categories, resulting in a sparse distribution of the aforementioned challenging factors. Consequently, there is a necessity to establish a dedicated environment that incorporates densely distributed challenging factors to facilitate robustness research. Compared with generic scenarios that are recorded by fixed or handheld cameras, visual tracking based on unmanned aerial vehicles (UAVs or drones) highlights challenges and requires more visual robustness. (1) *Tiny target*: the aerial overhead view causing the target size of a UAV-based system to be much smaller than other traditional datasets. (2) *Fast motion*: unlike fixed cameras, UAV-based datasets include both camera and target motion, resulting in frequent and drastic target position changes in consecutive frames. (3) *Abrupt variation*: due to the long distance between the target and UAV-mounted camera, a slight movement of UAV will lead to a drastic change in its viewpoint, making the visual information (both foreground and background) shift drastically between consecutive frames.

High-quality UAV-based benchmarks with the above challenging factors are critical to developing robust visual tracking algorithms. Although existing works have provided an important basis (Table 1), they still have several shortcomings:

- *Small-scale dataset* Early UAV datasets (Hsieh et al. 2017; Xia et al. 2018) usually cover only a few thousand images. Although recent works have improved the dataset scale, the size of any single task remains relatively small (Yu et al. 2020; Zhu et al. 2021; Bondi et al. 2020), often insufficient to support data-driven vision algorithms.
- *Scarcity of UAV-based data* Most UAV datasets (Xia et al. 2018; Mueller et al. 2016; Li and Yeung 2017; Bondi et al. 2020) contain multiple data sources, such as data collection from websites or data generated from the UAV simulators, but lack UAV data collected in real scenarios.
- *Limited UAV types* UAVs can be classified into fixed-wing, rotary-wind, and flapping-wing vehicles. Among the three, bionic UAVs with flapping-wing structure remains under exploration. However, the existing UAV datasets (Hsieh et al. 2017; Xia et al. 2018; Mueller et al. 2016; Li and Yeung 2017; Yu et al. 2020; Zhu et al. 2021; Bondi et al. 2020) all use fixed-wing or rotary-wing UAVs for data collection and lack attention to visual data from the flapping-wing UAVs.

The above problems motivate us to focus on new challenges posed by the aerodynamic structure of flapping-wing drones. Using the Large Wingspan bionic flight platform, a flapping-wing aircraft with cutting-edge flight performance made by our team, we construct the first **bio**nic **drone**-based visual benchmark **BioDrone** for SOT task. We summarize the characteristics of our benchmark and our contributions as follows.

- *Large-scale and high-quality benchmark with robust vision challenges* We take robust vision research as the entry point to construct BioDrone, which includes *600* videos with *304,209* manually labeled frames, and is annotated and reviewed under a precise process. To our knowledge, BioDrone is the first SOT benchmark collected by the bionic-based vision system and the largest UAV-based SOT benchmark. Figure 1 qualitatively compares BioDrone to other SOT benchmarks, demonstrating the impact of challenging factors on tracking performance. Most SOTA methods can maintain robust tracking for thousands of frames on generic benchmarks, but easily lose target after tens of frames on

Table 1 Summary of existing UAV-based datasets and generic SOT datasets ($1k=10^3$, $1m=10^6$)

Name	Year	Task	#Frames	# Videos	Resolution	Collection Way	
						UAV-based	Other Sources
CARPK (Hsieh et al. 2017)	2017	DET	$1.4k$	–	1280*720	Rotary-wing UAV (DJI Phantom 3 Professional)	N
DOTA (Xia et al. 2018)	2018	DET	$2.8k$	–	Various	Rotary-wing UAV	Multiple platforms (e.g., Google Earth)
UAV123 (Mueller et al. 2016)	2016	SOT	$110k$	123	1280*720	Rotary-wing UAV (DJI S1000)	UAV simulator (UE4)
UAV20L (Mueller et al. 2016)	2016	SOT	$58.7k$	20	1280*720	Rotary-wing UAV (DJI S1000)	UAV simulator (UE4)
DTB70 (Li and Yeung 2017)	2017	SOT	$15.8k$	70	1280*720	Rotary-wing UAV (DJI Phantom 2 Vision)	Website
UAVDT (Yu et al. 2020)	2018	VID, SOT, MOT	$80k$	100	1024*540	Rotary-wing UAV (DJI Inspire 2)	N
VisDrone (Zhu et al. 2021)	2018	DET, VID, SOT, MOT	$179k+10k$	263	3840*2160	Rotary-wing UAV (DJI Mavic and Phantom series)	N
BIRDSAI (Bondi et al. 2020)	2020	VID, MOT	$162k$	172	640*480	Fixed-wing UAV	UAV simulator (AirSim-W)
BioDrone	**2022**	**SOT**	**$304k$**	**600**	**1440*1080**	**Flapping-wing UAV**	**N**
OTB50 (Wu et al. 2013)	2013	SOT	$29k$	59	Various	N	Website
OTB100 (Wu et al. 2015)	2015	SOT	$59k$	100	Various	N	Website
VOT2016 (Kristan et al. 2016)	2016	SOT	$21.5k$	60	Various	N	Website
VOT2017 (Kristan et al. 2017)	2017	SOT	$21.3k$	60	Various	N	Website
TrackingNet (Muller et al. 2018)	2018	SOT	$14.4m$	$30.6k$	Various	N	Website
LaSOT (Fan et al. 2021)	2020	SOT	$3.87m$	$1.55k$	Various	N	Website
GOT-10k (Huang et al. 2021)	2021	SOT	$1.45m$	$10k$	Various	N	Website
VideoCube (Hu et al. 2023)	2022	SOT	$7.46m$	500	Various	N	Website

To our knowledge, BioDrone is the first SOT benchmark collected by the bionic-based vision system and the largest UAV-based SOT benchmark

Fig. 1 This paper aims to study the robust vision problem in visual object tracking; thus, we propose a bionic drone-based SOT benchmark named BioDrone to support this goal. In this figure, we compare BioDrone (**G**–**J**) with generic SOT benchmarks represented by VOT short-term tracking competition (Kristan et al. 2019b, a) (**A**, **B**), LaSOT (Fan et al. 2021) (**C**, **D**), VideoCube (Hu et al. 2023) (**E**, **F**). Here we select the same object categories (car and person) in different benchmarks, and add performances of state-of-the-art (SOTA) tracking methods for better comparison (green bounding-box represents ground-truth, yellow bounding-box represents KeepTrack (Mayer et al. 2021), blue bounding-box represents MixFormer (Cui et al. 2022), red bounding-box represents SiamRCNN Voigtlaender et al. 2020). Compared to other benchmarks, BioDrone highlights the challenges of *tiny target* and *fast motion*. The above factors can affect appearance and motion information, bringing troubles to most tracking algorithms on BioDrone. Most SOTA methods lose the target after tens of frames on BioDrone, but they perform well for thousands of frames on other benchmarks (Color figure online)

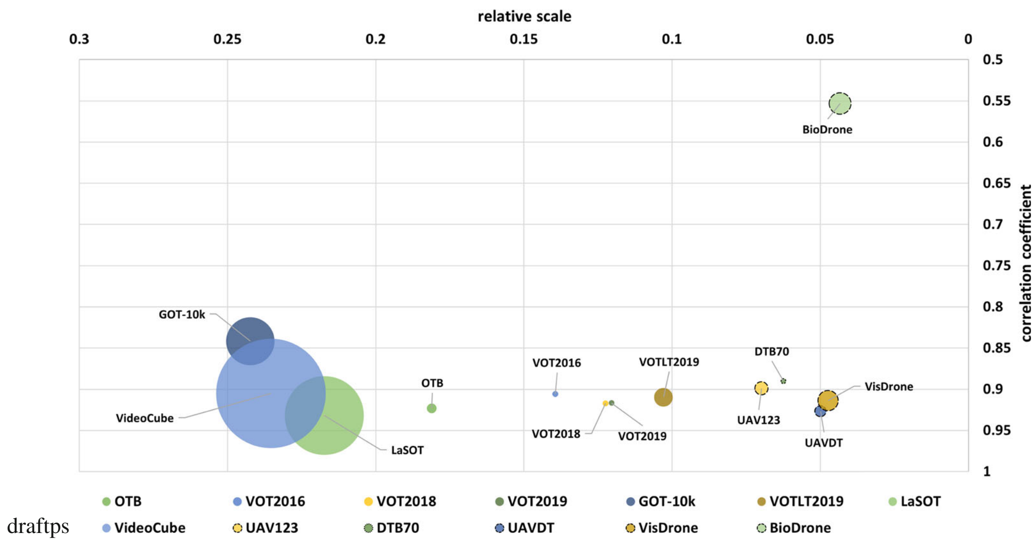


Fig. 2 Summary of existing SOT benchmarks, including classical benchmarks (OTB100 Wu et al. 2015, VOT2016 Kristan et al. 2016, VOT2018 Kristan et al. 2019b, VOT2019 Kristan et al. 2019a, GOT-10k Huang et al. 2021, VOTLT2019 Kristan et al. 2019a, LaSOT Fan et al. 2021, Videocube Hu et al. 2023), and UAV-based benchmarks (UAV123 Mueller et al. 2016, UAVDT Yu et al. 2020, DTB70 Li and Yeung 2017, VisDrone Zhu et al. 2021). The bubble diameter is in proportion to the total frames of a benchmark. The bubbles with dashed borders represent UAV-based benchmarks. The horizontal coordinate represents the average relative scale of the target, and the vertical coordinate represents the average correlation coefficient between consecutive frames. The proposed BioDrone has a *smaller target size* and *more drastic frame changes* between consecutive frames, with higher demands on the robustness of tracking algorithms

BioDrone. Figure 2 quantitatively compares BioDrone with others and indicates that *smaller target size* and *more drastic frame changes* between consecutive frames in BioDrone put higher demands on tracking robustness.

- *Videos from Bionic-based UAV* Unlike the existing UAV-based datasets that ignore the flapping-wing UAV structure, our team designs the Large Wingspan bionic flight platform with cutting-edge performance for data collection. Compared with other mechanical structures, the flapping-wing system has broader application prospects due to its lifelike bionic structure. Besides, the flapping-wing design includes additional visual challenges due to more damaging camera shake during the air movements, as shown in Fig. 3.
- *Rich challenging factor annotation* Different from existing UAV-based datasets (Mueller et al. 2016; Li and Yeung 2017; Yu et al. 2020; Zhu et al. 2021) that only provide sequence-level annotation for several challenging factors, BioDrone first provides high-quality fine-grained manual annotations (bounding-box and *occlusion* annotation) and automatically generate frame-level labels for ten challenge attributes, aiming to provide detailed information for further analyses.
- *Effective tracking baseline* As shown in Fig. 1, challenging factors in BioDrone cause algorithms to fail easily. Thus, we optimize the SOTA method KeepTrack (Mayer et al. 2021) and design a new baseline UAV-KT. Besides, we propose a suitable training strategy, and finally achieve a 5% performance boost in the precision score.
- *Comprehensive experimental analyses* BioDrone contains a complete evaluation mechanism and metrics, compares 20 represent methods and 3 proposed baselines, and analyzes their tracking performance in multiple dimensions, aiming to systematically explore the problems of robust vision brought by flapping-wing UAVs.

2 Related Work

2.1 Generic SOT Datasets

SOT (Wu et al. 2015) is a *category-independent* task, which intends to track a moving target without any assumption about the target category. This characteristic allows SOT to be suitable for open-set testing with broad prospects. Since 2013, several generic SOT datasets have been released to support related research.

As one of the earliest benchmarks, *OTB50*[1] (Wu et al. 2013) released in 2013 can be regarded as the earliest SOT benchmark for scientific evaluation. Two years later, *OTB100* (Wu et al. 2015) expands the original version for more comprehensive comparisons. Subsequently, the *VOT com-*

[1] http://cvlab.hanyang.ac.kr/tracker_benchmark/index.html.

(a) (b) (c)

Fig. 3 Example of typical UAVs. Compared to the other two types of UAVs, flapping-wing UAVs include more challenges due to their bionic mechanical structure. **a** Fixed-wing UAV (Hu et al. 2022). **b** Rotary-wing UAV (Müller et al. 2018). **c** Flapping-wing UAV.

petition[2] (Kristan et al. 2013, 2014, 2015, 2016, 2017, 2019b, a, 2020, 2021) series provide diverse and high-quality datasets to challenge algorithms.

With the advancement of data-driven trackers, datasets with larger scales are demanded. *GOT-10k*[3] (Huang et al. 2021) is a significant high-diversity short-term tracking dataset that comprises 10,000 videos with *one-shot protocol*. Long-term tracking dataset *LaSOT*[4] (Fan et al. 2021) has 3.8*m* manually labeled frames with 1,550 videos. It follows the one-shot protocol as well for improving tracking generalization. Recently, the global instance tracking dataset *VideoCube*[5] (Hu et al. 2023) is proposed to provide videos with shot-cut and scene-switching. Compared with other SOT datasets, VideoCube not only models the real world comprehensively but also challenges both the perceptual and cognitive components of trackers.

However, most sequences in these generic benchmarks are collected by fixed cameras, in which the target usually moves smoothly with a notable appearance. The distribution of challenging factors is sparse and usually requires data mining to support robust vision research.

2.2 UAV and UAV Vision

In 1879, French engineer Alphonse Pénaud created a rubber-band-powered aircraft to model the flapping-wing structure, which has been used for toy design due to its straightforward structure. However, restricted by technology, the research on flapping wing aircraft has progressed slowly. At this stage, the Wright brothers invented plane in 1903, and Paul Kearney prompted helicopter in 1907, causing fixed-wing and rotary-wing aircraft to occupy the sky, and promoting a series of research in the following decades (McMasters and Cummings 2002, 2004; Sims and Uhlig 1991). Recently, with the development of microcomputers, electrical engineering, and artificial intelligence, UAVs have gradually been favored worldwide, and significantly shortened the gap between enthusiasts and traditional large aircraft. UAVs are typically battery-powered, hand-launched, and belly-landed, and can be divided into three types: fixed-wing, rotary-wing, and flapping-wing, as shown in Fig. 3.

The appearances of the first two UAVs are similar to airplanes or helicopters, relying on fixed or rotating wings to provide power for their fuselages, and have been widely used by academia and industry applications, such as intelligent transportation, agricultural procedures, material conveyance, security surveillance, etc.(Fraire et al. 2015; Barrientos et al. 2010). Although the research of fixed-wing and rotary-wing UAVs has become increasingly sophisticated, their structures' shortcomings are also gradually explored. Defects like large size, insufficient mobility energy, and low efficiency motivate researchers to reconsider designing flapping-wing UAVs—a kind of bionic aircraft with high lift coefficient and flexible maneuverability for various task situations (Lee et al. 2018; Zhang and Rossi 2017). In recent years, flapping-wing UAVs have attracted growing attention due to their flexibility. It is worth noting prosperous information obtained by visual sensors installs a pair of "eyes" for the flapping-wing UAVs, enabling a prerequisite for accomplishing various tasks smoothly. This section will introduce representative flapping-wing UAVs and their vision system.

In 1988, researchers proposed the first flapping-wing UAV Microbat, which has a 15–20 cm wingspan and 20–30 Hz flapping frequency (Pornsin-Sirirak et al. 2001). In the same year, another flapping-wing UAV, Entomopter, is designed for Mars exploration (Rigelsford 2004). In 2016, DelFly II (De Croon et al. 2016), which contains an airborne stereoscopic perception system (two cameras that can collect visual images simultaneously at 30 Hz), was published for research. Flight experiments illustrate that it can successfully detect and avoid walls, but the short battery life and the poor imaging quality (720 × 240 resolution) restrict its application. Some other researchers modified a commercial flapping-wing UAV and equipped it with a lightweight first-person view (FPV) camera to realize the basic object tracking function (Ryu et

[2] https://votchallenge.net/.

[3] http://got-10k.aitestunion.com.

[4] https://cis.temple.edu/lasot/.

[5] http://videocube.aitestunion.com.

al. 2016). It has a vision algorithm integration system to communicate with the ground control system, which can transfer the captured images to the ground station in real time. However, the transmission system has a short communication distance, making it difficult to achieve long-distance tracking. Recently, another research group has developed Dove (Yang et al. 2018), which can transmit color video to the ground station. But its function is mainly limited to aerial photography, and there is still a broad space for development.

Consequently, the visual systems of existing flapping-wing UAVs are all airborne; sensors are mounted on the fuselage and provide environmental information like birds' eyes. However, specific defects like imaging quality and flight endurance limit the captured visual information. Therefore, although existing research on flapping-wing UAVs has been boosted, it is still difficult to construct high-quality visual datasets like fixed-wing or rotary-wing UAVs.

2.3 UAV-Based Tasks and Datasets

Encouraged by the eye-catching development of UAV-based research, various visual tasks have been applied in UAV systems to process environmental information. Since detection and tracking are closely related to UAV vision systems, most UAV-based datasets are constructed to support these two tasks.

Object detection (Liu et al. 2020) aims to accurately determine the category and location of targets, which can be further divided into image object detection (DET Russakovsky et al. 2015) and video object detection (VID Han et al. 2021). It's worth noting that the target category of object detection is generally restricted to pre-defined classes. *Car Parking Lot (CARPK)*[6] (Hsieh et al. 2017) is the first large-scale vehicle detection and counting dataset, which is collected by rotary-wing UAVs and covers nearly 90,000 cars in various parking lots. *DOTA*[7] (Xia et al. 2018) is another large-scale DET dataset with image resolution ranges from 800×800 to $20,000 \times 20,000$ pixels.

Object tracking (Zhu et al. 2021; Wu et al. 2021) can be further divided into single object tracking (SOT Hu et al. 2023; Fan et al. 2021; Huang et al. 2021) and multi-object tracking (MOT Luiten et al. 2021; Dendorfer et al. 2021). MOT usually combines with the VID task—algorithms should detect objects in the first frame, then calculate the similarity to determine instances with the same ID in consecutive frames. Conversely, SOT is a *category-independent* task, which intends to track a moving target without any assumption about the target category. *UAV123* and *UAV20L*[8] (Mueller et al. 2016) are pioneering works that construct UAV-based SOT datasets from three systems: a rotary-wing UAV, a low-cost UAV, and a UAV simulator (UE4[9]). Significant deviation (e.g., target scale and ratio) challenges classical SOT methods and invokes the following research in UAV-based visual tracking. *Drone Tracking Benchmark (DTB70)*[10] (Li and Yeung 2017) includes 70 video sequences to support short-term and long-term tracking. Some sequences are captured by a rotary-wing UAV, while others are collected from YouTube.

Besides, some other UAV datasets are designed to support multiple visual tasks. *UAV Detection and Tracking (UAVDT)*[11] (Yu et al. 2020) is a large-scale vehicle detection and tracking dataset, which includes 100 video sequences collected by rotary-wing UAVs to support multiple vision tasks like VID, SOT and MOT. *VisDrone*[12] (Zhu et al. 2021) combines 263 video clips with 179k frames and additional 10k static images to support DET, VID, SOT, and MOT. Recently, a challenging object detection and tracking dataset*BIRDSAI*[13] (Bondi et al. 2020) is published. As a multi-modality dataset, it includes 48 real videos collected by a TIR camera mounted on a fixed-wing UAV and 124 synthetic aerial TIR videos generated from AirSim-W simulator (Bondi et al. 2018).

Table 1 summarizes the existing generic and UAV-based SOT datasets. Most datasets are collected from websites or simulators, while the limited UAV data comes from rotary-wing or fixed-wing UAVs, lacking visual datasets collected by flapping-wing UAVs. This blank area motivates us to conduct this work and build the first bionic drone-based SOT benchmark to better support robust vision research.

3 BioDrone Benchmark

A high-quality benchmark labels the target in the video frame and provides criteria for algorithm evaluation. Particularly, benchmarks incorporating multiple challenging factors are critical for training and testing robust trackers.

As summarized in Sect. 2, existing benchmarks all ignore collecting data from bionic-based aircraft, motivating us to conduct BioDrone for robust vision research. BioDrone is collected by a state-of-the-art (SOTA) flapping-wing UAV and annotated under a precise process. It includes *600* videos with *304,209* manually labeled frames. The sequence length varies from *300* to *990* frames, and the average length is around *507*. To our knowledge, BioDrone is the first

6 https://lafi.github.io/LPN/.

7 https://captain-whu.github.io/DOTA/.

8 https://cemse.kaust.edu.sa/ivul/uav123.

9 https://www.unrealengine.com.

10 https://github.com/flyers/drone-tracking.

11 https://sites.google.com/site/daviddo0323/projects/uavdt.

12 https://github.com/VisDrone/VisDrone-Dataset.

13 https://sites.google.com/view/elizabethbondi/dataset.

SOT benchmark collected by a bionic-based aircraft and the largest UAV-based SOT benchmark.

3.1 Data Collection and Annotation

3.1.1 Data Collection

We use the Large Wingspan bionic flight platform for data acquisition. It is designed with a high degree of biological similarity in appearance and sporty performance, as shown in Fig. 4a. Compared with existing flapping-wing UAVs, Large Wingspan adopts a rotor-flapping composite power arrangement with a single-section wing streamlined aerodynamic layout. Its fuselage length is 800 mm, wingspan is 1500 mm, biplane flutter frequency is 0–4 Hz, and flight altitude is 5–100 m. Functional loads such as high-definition map transmission and network communication are also deployed in Large Wingspan, ensuring that it can collect visual images from higher altitudes.

In the data acquisition process, we set different flight attitudes for various scenes under three lighting conditions, ensuring that the raw data can fully reflect the robust visual challenges of the flapping-wing UAVs. In the original date processing process, no post-processing such as frame selection or editing was applied to the collected videos. Therefore, the sequences in the dataset are transformed from real-time recorded videos (30FPS), maintaining a consistent sample rate of 30 Hz.

3.1.2 Data Annotation and Quality Control

An experienced team precisely labels BioDrone by following two main rules: (1) using the tightest bounding-box to mark the visible part of the user-specified target; and (2) adding an absent label for out-of-view or full-occluded target. A strict three-round review process is executed to ensure the annotation quality. Experienced annotators are trained to conduct the preliminary work and self-inspection, then submit the result to verifiers for second-round verification. Finally, the authors judge whether to accept it in the third-round validation. Any rejection in the above processes will result in the re-annotation to guarantee a high-quality benchmark. The representative data of BioDrone is shown in Fig. 4b.

3.1.3 Subset Division

We divide BioDrone into the training set (300 videos), the validation set (100 videos), and the test set (200 videos). The sequence length distribution is illustrated in Fig. 5a; we ensure that the distribution on the three subsets is essentially the same. In particular, three representative algorithms (i.e., KeepTrack Mayer et al. 2021, MixFormer Cui et al. 2022, and SiamRCNN Voigtlaender et al. 2020) are selected to test the 600 videos, and the mean performance of the three trackers is regarded as the score of each sequence. We then organize 600 sequences according to their scores, and finally obtain the difficulty ranking of all data. The distribution of sequence difficulty in each subset is roughly the same. As shown in Fig. 5b, BioDrone includes three illumination conditions: bright light (295 videos), low light in the evening (241 videos), and low light at night (64 videos). Figure 5c indicates that BioDrone has two main target categories: person (295 videos) and vehicle (305 videos).

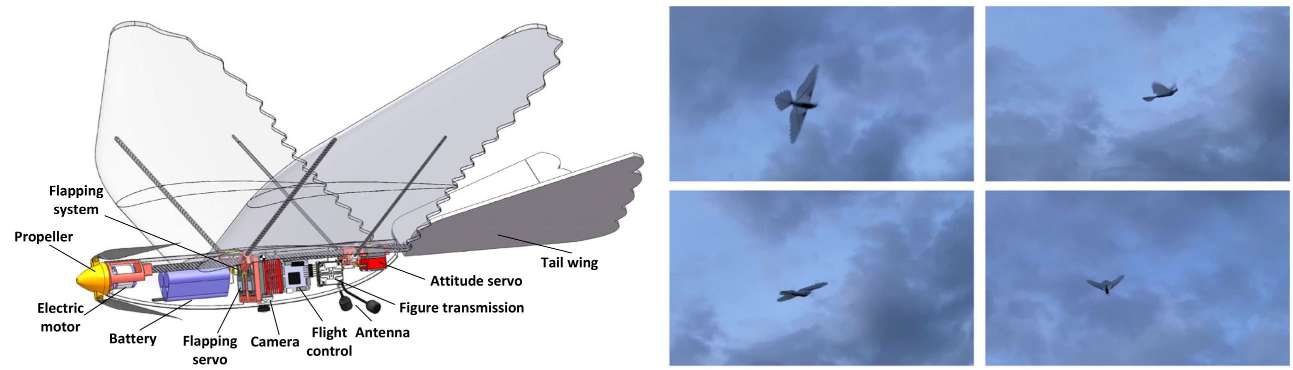
3.2 Challenging Attributes

The need for robust vision in SOT tasks is primarily from a large number of challenging factors in the environment. Notably, special collection situations (e.g., lens shake, the unique viewpoint, and the long shooting distance) bring more challenging factors to UAV-based datasets and require more robust algorithms to accomplish tracking tasks. However, we note that existing UAV-based datasets (Mueller et al. 2016; Li and Yeung 2017; Yu et al. 2020; Zhu et al. 2021) only provide sequence-level annotation for several challenging attributes—these coarse-grained labels cannot effectively provide detailed information for further analyses.

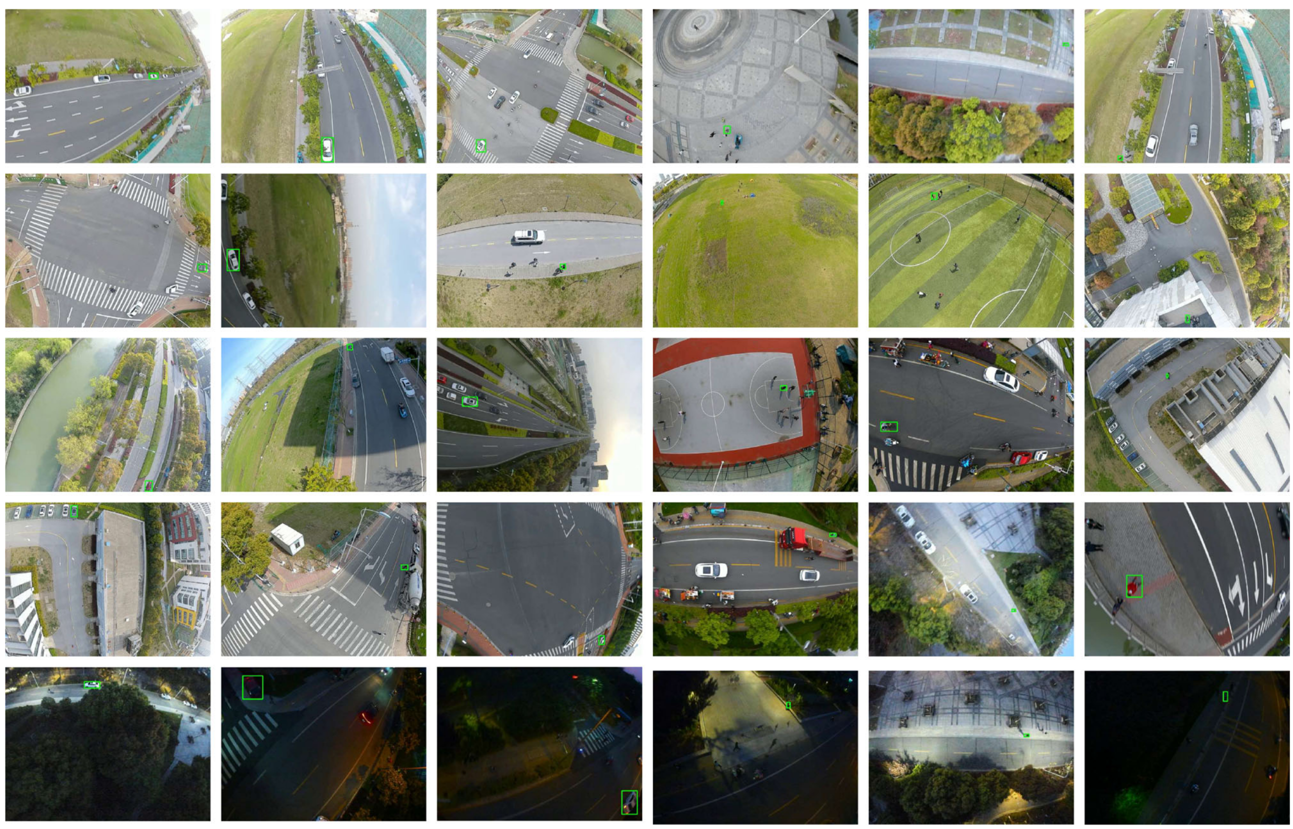
Therefore, we first provide high-quality frame-by-frame manual annotations (bounding-box and *occlusion* annotation) and automatically generate frame-level labels for ten challenge attributes based on SOTVerse (Hu et al. 2022) and VideoCube (Hu et al. 2023).

For the t-th frame F_t in a sequence $s_i = \{F_1, F_2, \ldots, F_t, \ldots\}$, BioDrone uses (x_t, y_t, w_t, h_t) (i.e., the coordinate information of the upper left corner and the shape of the bounding-box) like most classical benchmarks to represent the target bounding-box. Challenging attributes in BioDrone are two categories: *static attributes* only relate to the current frame, while *dynamic attributes* record changes between consecutive frames. The calculation rules for static attributes are as follows:

- *Target aspect ratio and scale* Target *ratio* is defined as $r_t = h_t/w_t$, and target scale is calculated via $s_t = \sqrt{w_t h_t}$. Specifically, we calculate *relative scale* by $s_t' = s_t/\sqrt{W_t H_t}$ to eliminate the influence of image resolution, where W_t and H_t represent the image resolution of F_t.
- *Illumination condition* Visual information recorded in special light conditions can be transferred to standard illumination by multiplying a correction matrix C_t (Finlayson and Trezzi 2004). Thus, BioDrone quantifies the *illumination* by calculating the Euclidean distance between C_t and $\mathbf{1}^{1\times 3}$.
- *Image clarity* BioDrone uses the *blur box* degree to measure the image clarity, which is generated by Laplacian



(a) Schematic diagram of Large Wingspan bionic flight platform and its flight attitudes.



(b) The representative data of BioDrone. Each video is strictly collected based on duration, instance classes, main scene categories, and illumination.

Fig. 4 Illustrations of the flapping-wing UAV used for data collection and the representative data of BioDrone. Different flight attitudes for various scenes under three lighting conditions are included in the data acquisition process, ensuring that BioDrone can fully reflect the robust visual challenges of the flapping-wing UAVs

transform (Pech-Pacheco et al. 2000). We convert the RGB bounding-box into gray-scale G_t, then convolve G_t with a Laplacian kernel, and calculate the variance as clarity.

Several dynamic attributes can be directly calculated from static attributes. Correspondingly, the variations of the above static attributes in two sequential frames are defined as *delta ratio*, *delta relative scale*, *delta illumination* and *delta blur box*. Besides, BioDrone also supplies another two dynamic attributes for in-depth analyses:

- *Target motion Fast motion* is selected to quantify the target center distance between consecutive frames by

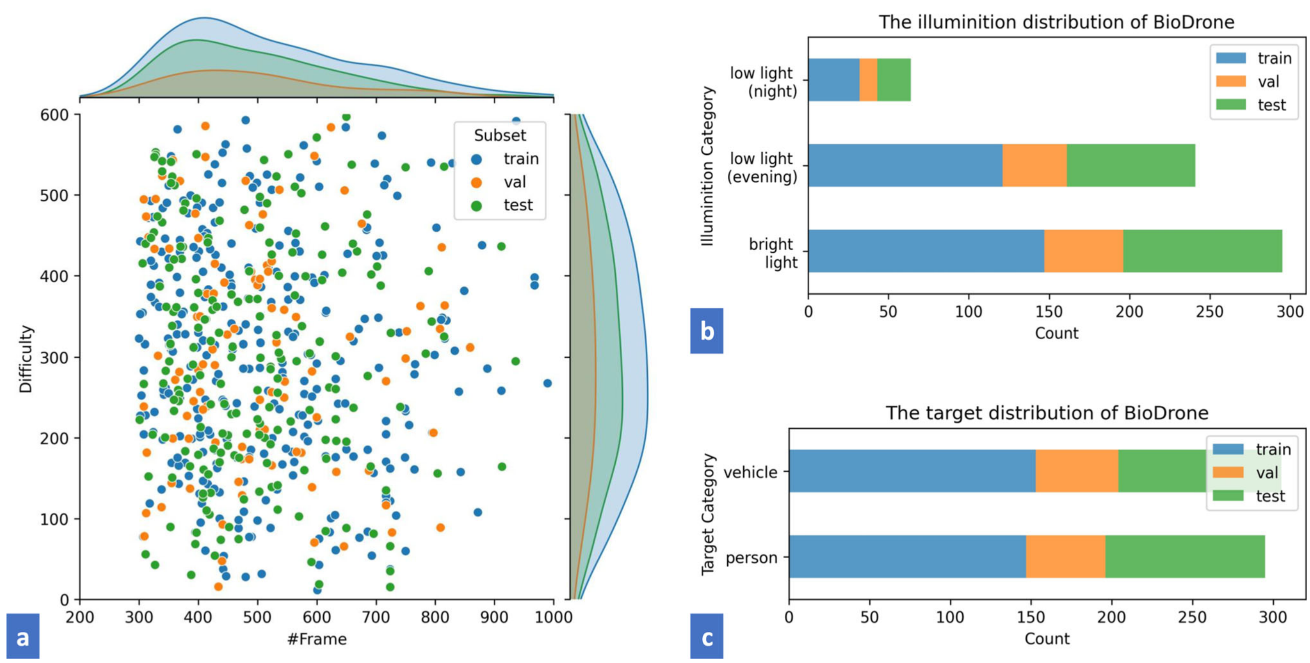


Fig. 5 Data distribution of BioDrone. The data distribution of different dimensions keeps consistent in each subset. **a** The distribution of sequence lengths and tracking difficulties. **b** The distribution of illumination conditions. **c** The distribution of target categories

$d_t = \|c_t - c_{t-1}\|_2 / max(s_t, s_{t-1})$. Note that we do not distinguish between the specific causes of the target center distance (e.g., target motion or camera motion), but rather focus on the disruption of the target trajectory due to fast motion. For instance, some SOT algorithms only locate the target position in the next frame within a limited search region near the result of the previous frame. However, fast motion can disrupt the continuity of the target's motion trajectory (e.g., the target's position in the next frame is likely to exceed the search region of the algorithm) and challenge the tracking robustness.

- *Integrated variation between consecutive frames Correlation coefficient* is a metric used to measure the similarity between current frame F_t and the previous frame F_{t-1}. BioDrone selects the Pearson product-moment correlation coefficient $\rho_t = \frac{\text{cov}(F_t, F_{t-1})}{\sigma_{F_t}\sigma_{F_{t-1}}}$, in which the numerator calculates the covariance of F_t and F_{t-1}, and the denominator is the product of the standard deviation. The correlation coefficient reflects the changes between consecutive frames and has been normalized in [0, 1].

To further demonstrate the challenges of BioDrone, we compare the attribute distributions of BioDrone and other SOT benchmarks (frame-level annotations are provided by SOTVerse), then plot the attribute distributions in Fig. 6. Compared with other SOT benchmarks, BioDrone includes more *tiny targets* (Fig. 6a, b) with *more drastic variations* (Fig. 6c, d) between consecutive frames, which provides a high-quality test bed for further research.

4 Trackers

4.1 Single Object Tracking Methods

Table 2 shows 20 representing SOT algorithms covering both classic and SOTA methods. Here, we list the basic information about these trackers.

KCF (Henriques et al. 2014) is a classical correlation filter (CF) based method, which balances high speed and tracking accuracy, and becomes a representative tracking framework in the early days. ECO (Danelljan et al. 2017) combines convolutional neural networks (CNN) with CF, aiming to use deep networks to improve feature representation. The feature representation of ECO is a combination of the first and last convolutional layer in the VGG-m (Chatfield et al. 2014), along with histogram of oriented gradient (HOG) (Dalal and Triggs 2005) and color names (CN) (Van De Weijer et al. 2009).

As the originator of siamese neural network (SNN) based trackers, SiamFC (Bertinetto et al. 2016) achieves satisfactory tracking performance by matching features between the template region and the search region through a simple network structure. It uses AlexNet (Krizhevsky et al. 2021) for feature representation and matches features via cross-correlation operation. After that, SiamRPN (Li et al. 2018) select the region proposal network (Girshick 2015) to achieve accurate target regression, DaSiamRPN (Zhu et al. 2018) uses data augmentation to enhance the discriminative ability, SiamRPN++ (Li et al. 2019) and SiamDW (Zhang and Peng

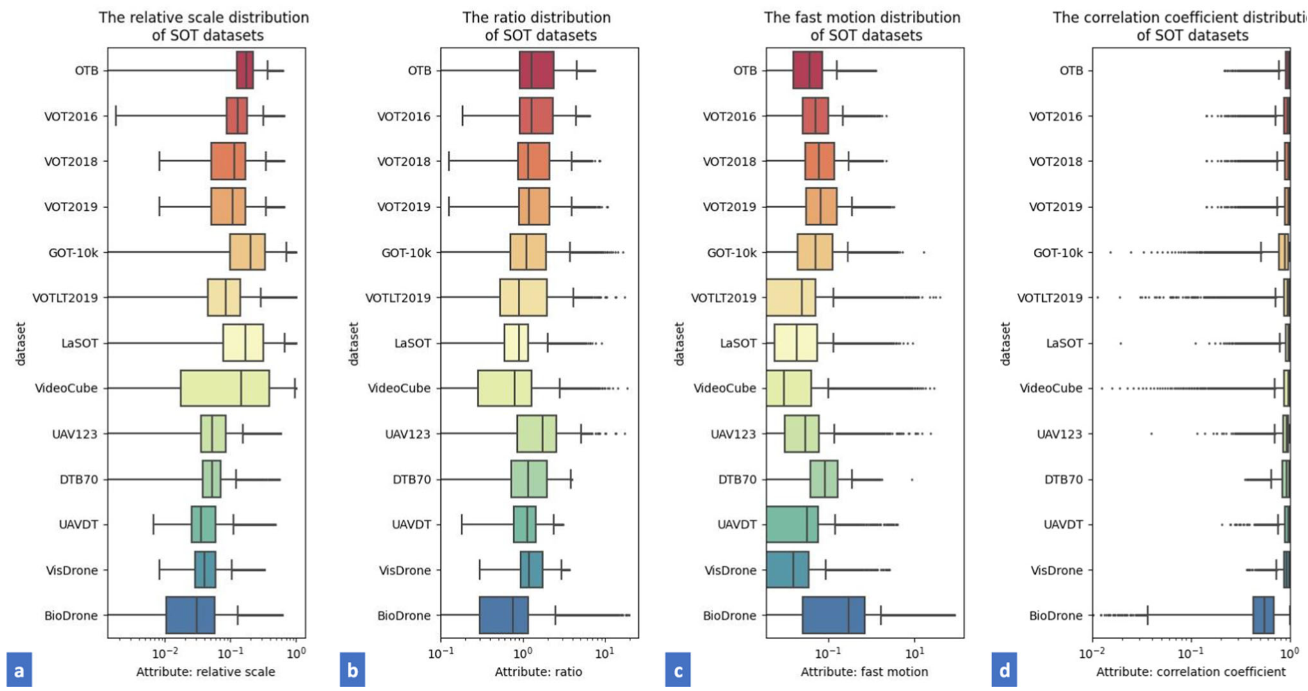


Fig. 6 Challenging attributes distribution of BioDrone and representative SOT benchmarks. **a** The distribution of relative scale (smaller value means including more tiny targets). **b** The distribution of aspect ratio (smaller or larger value means including more irregular shapes). **c** The distribution of fast motion (larger value means including faster target movement). **d** The distribution of correlation coefficient (smaller value means including more drastic variations between consecutive frames). Clearly, BioDrone includes more *tiny targets* with *more drastic variations* between consecutive frames, and requires more robust methods to accomplish target tracking

Table 2 Characteristic of the single object tracing methods in this work

Tracker	Publish	Feature Representation	Matching Operation	Update
KCF (Henriques et al. 2014)	TPAMI'15	HOG	Correlation Filter	Y
SiamFC (Bertinetto et al. 2016)	ECCV'16	AlexNet	Cross Correlation	
ECO (Danelljan et al. 2017)	CVPR'17	VGG-m	Correlation Filter	Y
SiamRPN (Li et al. 2018)	CVPR'18	AlexNet	Cross Correlation	
DaSiamRPN (Zhu et al. 2018)	ECCV'18	AlexNet	Cross Correlation	
ATOM (Danelljan et al. 2019)	CVPR'19	ResNet-18	Correlation Filter	Y
SiamRPN++ (Li et al. 2019)	CVPR'19	ResNet-50	Cross Correlation	
SiamDW (Zhang and Peng 2019)	CVPR'19	ResNet-22	Cross Correlation	
DiMP (Bhat et al. 2019)	ICCV'19	ResNet-50	Correlation Filter	Y
GlobalTrack (Huang et al. 2020)	AAAI'20	ResNet-50	Hadamard Correlation	
SiamFC++ (Xu et al. 2020)	AAAI'20	AlexNet	Cross Correlation	
Ocean (Zhang et al. 2020)	ECCV'20	ResNet-50	Cross Correlation	
KYS (Bhat et al. 2020)	ECCV'20	ResNet-50	Correlation Filter	Y
SiamCAR (Guo et al. 2020)	CVPR'20	ResNet-50	Cross Correlation	
PrDiMP (Danelljan et al. 2020)	CVPR'20	ResNet-50	Correlation Filter	Y
SuperDiMP (Danelljan et al. 2020)	CVPR'20	ResNet-50	Correlation Filter	Y
SiamRCNN (Voigtlaender et al. 2020)	CVPR'20	ResNet-101	Concatenate and Re-detection	Y
KeepTrack (Mayer et al. 2021)	ICCV'21	ResNet-50	Correlation Filter	Y
TCTrack (Cao et al. 2022)	CVPR'22	Temporally Adaptive CNN	Adaptive Temporal Transformer	Y
MixFormer (Cui et al. 2022)	CVPR'22	Mixed Attention Module		Y

CNN Convolutional Neural Network, *HOG* Histogram of Oriented Gradient

2019) introduce deeper and wider backbones (ResNet He et al. 2016) for feature extraction. Besides the development of backbone utilization, SiamFC++ (Xu et al. 2020), Ocean (Zhang et al. 2020), and SiamCAR (Guo et al. 2020) employ an anchor-free structure (Tian et al. 2019) to eliminate the dependence on anchors. Recently, SiamRCNN (Voigtlaender et al. 2020) utilizes a re-detection mechanism (based on FasterRCNN Ren et al. 2015) and proposes a tracklet dynamic programming algorithm to process object disappearance.

Another series of works started by ATOM (Danelljan et al. 2019) tries to combine CF and SNN together, and proposes a new framework to combine offline training and online updating. Based on the framework, DiMP (Bhat et al. 2019) optimizes the loss function for stronger discriminative ability, PrDiMP and SuperDiMP (Danelljan et al. 2020) use probabilistic regression to improve the accuracy. KeepTrack (Mayer et al. 2021) combines SuperDiMP with a target candidate association network, which is re-trained on hard sequences mined from LaSOT (Fan et al. 2021).

Some other works design custom networks to solve specific problems like target absence or similar instance interference. GlobalTrack (Huang et al. 2020) aims to keep tracking performance in long sequences; it does not assume motion consistency and performs a full-image search to eliminate cumulative error. KYS (Bhat et al. 2020) aims to better use scene information in the tracking process; it represents scene information as state vectors and combines them with the appearance model to locate the object. TcTrack (Cao et al. 2022) and MixFormer (Cui et al. 2022) are the two newest methods based on the transformer structure. TcTrack (Cao et al. 2022) is designed for object tracking in UAV-based scenes, which aims to fully exploit temporal contexts for aerial tracking. MixFormer (Cui et al. 2022) designs an end-to-end transformer-based framework to simultaneously accomplish feature extraction and target information integration.

4.2 New Baselines

As we analyzed in Sect. 1, challenging factors such as *tiny target* and *fast motion* cause algorithms to lose the target easily. Although some methods have combined a re-detection mechanism, fast motion makes it difficult to relocate the target via continuous trajectories, while the small object size significantly limits available appearance information. Thus, it is easy for trackers to relocate interferers rather than the target. Based on the above analyses, we optimize the SOTA method KeepTrack (Mayer et al. 2021), which employs a learned target candidate association network to track both the target and distractor objects, and design a new baseline UAV-KT for BioDrone (Fig. 7).

4.2.1 Base Model: KeepTrack

To improve the robust tracking ability when facing similar object interference, KeepTrack (Mayer et al. 2021) designs a mechanism to keep track of distractor objects. It chooses SuperDiMP (Danelljan et al. 2020) as the baseline, and adds a learnable correlation network to propagate the identity of all candidate targets in the tracking process. KeepTrack contains a classification branch and a bounding-box regression branch. The classification branch first obtains the score map through the SuperDiMP network, then generates the coordinates of candidates by selecting points that satisfy the requirements (i.e., the score is a local maximum and should exceed the threshold). Afterward, candidates' features are extracted and sent to the target candidate association network for candidate matching and location information generation. The regression branch follows the IoUNet (Jiang et al. 2018) utilized in ATOM (Danelljan et al. 2019) to precisely regress the bounding-box, and the target position information obtained from the classification branch is used to obtain and refine its position. Please refer to the original paper for more detailed information on the above two branches. Since our improvements are mainly concentrated in the candidate target matching network, here we briefly describe its structure in KeepTrack as follows.

Problem formulation KeepTrack defines the set of target candidates corresponding to the previous frame and the current frame, including distractors and targets, as $V^{'}$ and V. $V = \{v_i\}_i^N$, where N denotes the number of candidates appearing in each frame. The target candidate association problem for two subsequent frames is also formulated as finding the assignment matrix A between the two sets $V^{'}$ and V.

Target candidate extraction KeepTrack first processes the score map by selecting points that meet the requirements as candidate locations and extracts their features. After that, KeepTrack uses the candidate location c_i as a strong cue, then selects the candidate score $s(c_i)$ and the feature $f_i = f(c_i)$ obtained after a learnable convolutional layer as the other two complementary cues. Finally, a feature tuple is created for each candidate and is combined in the following way:

$$z_i = f_i + \varphi(s_i, c_i), \forall v_i \in V \tag{1}$$

where φ denotes a multilayer perceptron that maps s and c to the same dimensional space as f_i.

Candidate embedding network To get more representative candidate features, KeepTrack uses sparse feature matching to exchange z_i with bilateral information and self-information. Finally, a new more robust feature representation h_i is obtained.

Candidate matching The similarity matrix S, which is obtained by the dot product operation of $S_{i,j} =< h_i^{'}, h_j >$,

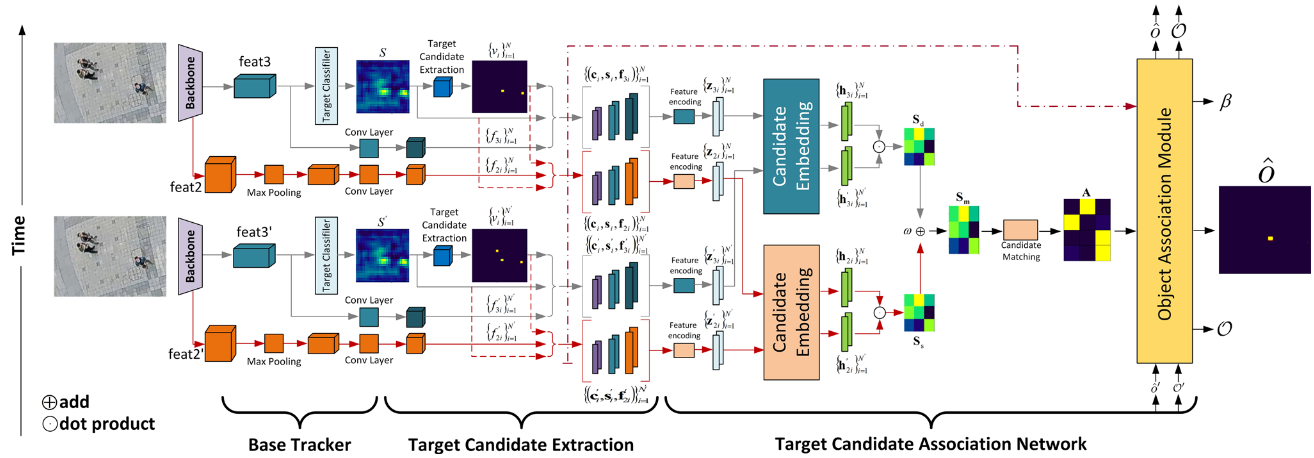


Fig. 7 Overview structure of the proposed new baseline UAV-KT based on KeepTrack (Mayer et al. 2021). The parts connected by red arrows represent our proposed shallow target candidate feature association network module, including target candidate feature extraction, production, embedding, and other operations. The parts connected by gray arrows are the original modules of KeepTrack. The score matrices obtained from different depth features are summed by a learnable coefficient w and perform matching and association operations (the parts connected by black arrows). Since the improvements are closely related and parallel to the original structure of KeepTrack, we draw UAV-KT based on KeepTrack to show the similarities and differences between these two methods clearly

is used to represent the similarity of candidates in $V^{'}$ and V. Due to situations like occlusion, disappearance, new appearance, or reappearance, the candidate targets do not necessarily have a definite correspondence within $V^{'}$ and V. However, the candidates must have a definite correspondence result to support the following process. Therefore, KeepTrack designs a dustbin to match candidates without correspondence (DeTone et al. 2018; Sarlin et al. 2020). Finally, an augmented assignment matrix A is obtained, in which an additional row and column are added to represent the dustbin. Note that a dustbin is a virtual candidate without any feature representation, and a candidate corresponds only if its similarity to all other candidates is low to a dustbin.

Object association A library O is used to keep track of each object that appears in the scene over time, in which each entry is an object that is visible in the current frame. When tracked online, the estimated assignment matrix A is used to determine the situation of objects (i.e., disappear, newly appear, or remain visible), and the visible objects can be explicitly associated and help in reasoning the target object $\hat{O}$.

Besides, KeepTrack also allows online updating. It describes a memory sample confidence score to decide whether to keep a sample in memory or not, and old samples will be replaced when a fixed memory size is used.

4.2.2 A New Baseline: UAV-KT

KeepTrack performs well among the representative SOT trackers in Sect. 4.1. However, due to the robust vision challenges introduced by the BioDrone benchmark, the original KeepTrack still has some limitations, motivating us to make appropriate modifications to obtain a more suitable model architecture.

Compared with generic object tracking, the tiny target in BioDrone not only lacks appearance information, but also needs wider receptive fields of deeper-level features to locate its position. On the one hand, deep features can obtain rich high-level semantic information, but cannot compensate for the lost pixel information for tiny targets. On the other hand, the smaller receptive field of low-level features can avoid the information loss problem, but it mainly extracts spatial information and ignores important semantic information (e.g., assumed as high-level features like temporal and spatial relationships, forward and backward scenes logical relationships, etc.). Based on the above analyses, a proper feature fusion module is added in KeepTrack to generate a new baseline named UAV-KT, which aims to improve the capability of tracking tiny targets in BioDrone.

Design of target candidate matching network based on different depth backbone features. As shown in Fig. 7, the red arrows represent operations of the new target candidate matching network proposed by UAV-KT, in which the feature map in the shallow block of the backbone is selected as a new cue, aiming to enhance the candidate target features and facilitate the ability of target candidate matching.

Unlike the original KeepTrack, we extend the target candidate matching network into two parallel networks for processing backbone features of different depths. The results of these processes are fused to obtain the final matching results. The operation on the shallow features and the information fusion method are described as follows:

Algorithm 1 Target candidate association algorithm

Input:
V: Set of target candidates;
$Z'_2(V_i)$: Set of embedded features of the previous frame;
S: Depth target candidate feature matching score matrix
Output:
$\hat{O}$: Target candidate association and matching module

$N = |V|$ `// Initialize`
for $i \leftarrow 1$ **to** N **do**
    `// Extraction via id`
    $f_2(V_i) \leftarrow$ extract from $feat_{backbone}$
        `// Extract backbone features`
    $f_2(V_i) \leftarrow \mathsf{MAXPOOL}(\mathsf{CONV}(f_2(V_i)))$
        `// Produce target candidate features`
    $z_2(V_i) = \mathsf{ADD}(\Phi(c(V_i) + s(V_i)), f_2(V_i))$
        `// Feature integration`
    $h_2(V_i), h'_2(V_i) \leftarrow \mathsf{EMBED}(z_2(V_i), z'_2(V_i))$
$S_s \leftarrow \{h_2(V_i)\}_{i=1}^N \odot \{h'_2(V_i)\}_{i=1}^N$
    `// Obtain score matrix`
$S_m = \mathsf{ADD}(w[0] * S_d, w[1] * S_s)$
    `// Fusion score matrix`
$\hat{O} \leftarrow$ match and associate by S_m
    `// Target candidate association and matching module`
return $\hat{O}$

- *Step 1* The shallow features $feat_2$ of the target candidates extracted from the backbone are fed into a maximum pooling layer and a learnable convolution layer to obtain a more discriminative appearance f_{2i} of the same size as f_{3i}.
- *Step 2* f_{2i} is encoded respectively with the target candidate coordinates and scores according to Eq. 1 to obtain the shallow target candidate features z_{2i}.
- *Step 3* The shallow target candidate features of the current frame and past frame are fed into the target candidate embedding network for information exchange and extraction, and finally generate richer and more robust features h_{2i}, h'_{2i}. The dot-product operation is performed on them to obtain the score map S_s.
- *Step 4* Here, the fusion operation is performed to obtain the final score matrix S_m. Notably, we introduce a learnable weight w to control the effect of different depth features, which is borrowed from BiFpn (Tan et al. 2020). The final score matrix is calculated by:

$$
\begin{aligned}
S_m &= w[0] * S_d + w[1] * S_s \\
w[i] &= \frac{w[i]}{\sum_{i=0}^{1} w[i] + \varepsilon}
\end{aligned}
\tag{2}
$$

  where $w[i]$ denotes the learnable weight set in the net, ε is a constant, generally set to 1×10^{-4}.
- *Step 5* Finally, the fused score matrix is used for subsequent operations such as candidate association and object association.

4.2.3 Training Strategies

Unlike large-scale general benchmarks, BioDrone is designed for robust vision research based on the flapping-wing UAV scenario, which contains multiple challenging factors. Therefore, a reasonable training strategy can help trackers enhance robustness in facing challenging factors such as tiny targets, fast motion, and interfering objects. In this section, we illustrate the detailed training strategies for the BioDrone benchmark and propose the re-trained baselines named KeepTrack* and UAV-KT*.

Generic SOT benchmarks include LaSOT (Fan et al. 2021), GoT-10k (Huang et al. 2021), and the proposed BioDrone are selected to re-train the base tracker (the left part in Fig. 7), which makes the tracker more robust in tracking tiny targets with fast motion in the UAV-based tasks. We sample multiple training and test frames from a video sequence to form training sub-sequences. 40k sub-sequences with a weight of 1:1:1 for each dataset are obtained for training the base tracker. The training and testing processes are conducted in a server with 4 NVIDIA TITAN RTX GPUs and a 64 Intel(R) Xeon(R) Gold 5218 CPU @ 2.30 GHz. We use adaptive moment estimation (Adam) with a batch size of 32 to train our model, in which the learning rate decay by 0.2 every 20th epoch with a learning rate of 2×10^{-4}. We train 30 epochs and freeze the first half of the weights of the backbone network during the training period.

The original KeepTrack and the proposed UAV-KT are trained based on the above training strategy to generate KeepTrack* and UAV-KT*. Furthermore, we notice that a proper training strategy is important—training different parts of the module (e.g., the target candidate association network) by BioDrone may decrease the performance of the original versions. Please refer to Sect. 5.3.2 for detailed results and analyses.

(a) One pass evaluation (OPE) mechanism by OTB benchmark [29].

(b) OPE system with re-initialization (R-OPE) mechanism by VideoCube benchmark [5].

Fig. 8 Execution process of two evaluation mechanisms. **a** The traditional OPE mechanism proposed by the OTB benchmark, in which the trackers keep tracking during the whole sequence. **b**) The R-OPE mechanism proposed by VideoCube, in which trackers will be re-initialized in the next frame when tracking failure (i.e., the IoU of predicted result p_t and ground-truth g_t $\frac{p_t \bigcap g_t}{p_t \bigcup g_t} < 0.5$) occurs

5 Evaluation and Experiments

5.1 Evaluation Protocol

5.1.1 Mechanisms

SOT tasks use two evaluation systems—OPE and the re-initialization mechanism (R-OPE). OPE mechanism initializes a tracker in the first frame and continuously records the results, which has been widely used by classical benchmarks (Wu et al. 2015; Fan et al. 2021; Huang et al. 2021). Recently, VideoCube (Hu et al. 2023) provides the R-OPE mechanism, which re-initializes the tracker when it fails in ten consecutive frames. BioDrone provides the above two mechanisms for performance evaluation, as shown in Fig. 8.

5.1.2 Metrics

For the t-th frame F_t in a sequence $s_i = \{F_1, F_2, \ldots, F_t, \ldots\}$, the positional relationship (e.g., intersection over union (IoU) and center distance) between predicted result p_t and ground-truth g_t is usually selected to calculate tracking performance. Like other SOT benchmarks, all evaluation indicators in BioDrone are based on the relationship between two bounding-boxes and their center points (i.e., the predicted center point c_p and the actual center point c_g). Note that target absent is regarded as an empty set (i.e., $g_t = \phi$).

Precision (PRE) Traditional *precision* score is calculated by:

$$d_c = \left\| c_p - c_g \right\|_2$$
$$\mathcal{P}(\theta_d) = \frac{1}{|\mathcal{G}|} \sum_{s_i \in \mathcal{G}} \frac{1}{|s_i|} |\{F_t : d_c \leq \theta_d\}|$$
$$P_{score} = \frac{1}{|\mathcal{G}|} \sum_{s_i \in \mathcal{G}} \frac{1}{|s_i|} |\{F_t : d_c \leq 20\}| \tag{3}$$

where $|\cdot|$ is the cardinality, θ_d is a threshold to judge whether the tracking result is precise. The precision score of s_i is defined as the proportion of frames whose center distance $d_c \leq \theta_d$. Calculating the mean value of each sequence s_i under video group $\mathcal{G}$ can generate the final precision score $\mathcal{P}(\mathcal{G})$. Previous works (Wu et al. 2015; Fan et al. 2021; Muller et al. 2018) usually draw the statistical results based on different θ_d into a curve named *precision plot*. Typically, $\theta_d = 20$ is widelyused to rank trackers (P_{score}).

Normalized precision (N-PRE) Recent work (Hu et al. 2023) indicates that the PRE score ignores the influence of the target scale, and provides a normalized precision score named N-PRE to solve this problem. Trackers with a predicted center outside the ground-truth rectangle will add a penalty item $d_c{}^p$ (i.e., the shortest distance between center point c_p and the ground-truth edge). For trackers whose center point falls into the ground-truth rectangle, the center distance $d_c{}'$ equals the original precision d_c (i.e., $d_c{}^p = 0$). Besides, to exclude the influence of target size and frame resolution, N-PRE selects the maximum value in frame F_t to normalize the result. The calculation can be summarized as:

$$\mathcal{N}(d_c{}') = \frac{d_c{}'}{\max(\{d_i{}' \mid i \in F_t\})}$$
$$\mathcal{P}'(\theta_d') = \frac{1}{|\mathcal{G}|} \sum_{s_i \in \mathcal{G}} \frac{1}{|s_i|} \left|\left\{F_t : \mathcal{N}(d_c{}') \leq \theta_d'\right\}\right|$$
$$P_{score}' = \frac{1}{|\mathcal{G}|} \sum_{s_i \in \mathcal{G}} \frac{1}{|s_i|} \left|\left\{F_t : c_p \in g_t\right\}\right| \tag{4}$$

Table 3 Performance of generic SOT trackers and the proposed baselines based on OPE and R-OPE mechanisms

Tracker	OPE Mechanism			R-OPE Mechanism				
	$P_{score}\uparrow$	$P'_{score}\uparrow$	$S_{score}\uparrow$	$P_{score}\uparrow$	$P'_{score}\uparrow$	$S_{score}\uparrow$	$L_{max}\uparrow$	$R_{count}\downarrow$
KCF (Henriques et al. 2014)	0.052	0.077	0.047	0.363	0.438	0.311	2.880	13.980
SiamFC (Bertinetto et al. 2016)	0.131	0.161	0.104	0.535	0.583	0.414	44.775	9.015
SiamDW (Zhang and Peng 2019)	0.151	0.210	0.126	0.550	0.626	0.434	50.340	8.640
Ocean (Zhang et al. 2020)	0.158	0.167	0.134	0.625	0.636	0.507	73.740	7.800
SiamFC++ (Xu et al. 2020)	0.162	0.180	0.139	0.568	0.594	0.482	71.175	8.595
DaSiamRPN (Zhu et al. 2018)	0.163	0.188	0.133	0.551	0.605	0.448	69.870	8.395
SiamRPN (Li et al. 2018)	0.173	0.199	0.139	0.557	0.606	0.448	69.115	8.245
SiamCAR (Guo et al. 2020)	0.213	0.235	0.178	0.655	0.672	0.530	94.000	6.480
TCTrack (Cao et al. 2022)	0.231	0.255	0.192	0.644	0.671	0.529	90.095	6.725
GlobalTrack (Huang et al. 2020)	0.237	0.249	0.183	0.560	0.570	0.451	55.515	6.865
ECO (Danelljan et al. 2017)	0.243	0.299	0.184	0.678	0.739	0.510	85.330	6.130
SiamRPN++ (Li et al. 2019)	0.315	0.337	0.241	0.685	0.703	0.528	116.620	5.030
ATOM (Danelljan et al. 2019)	0.341	0.385	0.285	0.754	0.787	0.623	141.420	4.180
DiMP (Bhat et al. 2019)	0.379	0.412	0.318	0.763	0.788	0.635	151.330	3.960
KYS (Bhat et al. 2020)	0.380	0.411	0.315	0.771	0.798	0.642	158.735	3.770
PrDiMP (Danelljan et al. 2020)	0.409	0.433	0.341	0.777	0.796	0.652	156.905	3.540
SuperDiMP (Danelljan et al. 2020)	0.426	0.447	0.361	0.784	0.796	0.658	163.785	3.565
MixFormer (Cui et al. 2022)	0.458	0.466	0.399	0.782	0.786	<u>0.675</u>	159.110	3.33
SiamRCNN (Voigtlaender et al. 2020)	0.468	0.474	0.394	0.720	0.726	0.616	119.335	4.455
KeepTrack (Mayer et al. 2021)	***0.504***	***0.523***	***0.424***	<u>0.803</u>	<u>0.817</u>	***0.673***	***170.000***	***3.075***
UAV-KT	<u>0.513 (0.009↑)</u>	<u>0.537 (0.014↑)</u>	<u>0.428 (0.003↑)</u>	***0.797***	***0.814***	0.663	<u>172.660</u>	<u>2.930 (0.145↓)</u>
KeepTrack*	*0.538 (0.034↑)*	*0.551 (0.028↑)*	*0.457 (0.033↑)*	**0.832**	**0.838**	**0.703**	**181.185**	*2.660 (0.415↓)*
UAV-KT*	**0.554 (0.050↑)**	**0.568 (0.045↑)**	**0.466 (0.041↑)**	*0.822*	*0.832*	*0.691*	*180.595*	**2.605 (0.470↓)**

The top-4 trackers are highlighted by bold, italic, underline, and bolditalic. Clearly, the proposed UAV-KT* baseline performs better in different evaluation mechanisms and metrics, demonstrating its robustness improvement compared to the baseline KeepTrack (Mayer et al. 2021)

Draw statistical results based on different $\theta_d^{'} \in [0, 1]$ into a curve generates the *normalized precision plot*. Particularly to overcome the influence of threshold selection, the proportion of frames whose predicted results successfully fall in the ground-truth rectangle is used to rank trackers ($P^{'}_{score}$).

Success. Like the calculation process in the precision plot, traditional *success* score of frame F_t is calculated by:

$$
\begin{aligned}
s_t &= \Omega(p_t, g_t) = \frac{p_t \bigcap g_t}{p_t \bigcup g_t} \\
\mathcal{S}(\theta_s) &= \frac{1}{|\mathcal{G}|} \sum_{s_i \in \mathcal{G}} \frac{1}{|s_i|} |\{F_t : s_t \leq \theta_s\}| \\
S_{score} &= \frac{1}{|\Theta_s|} \sum_{\theta_s \in \Theta_s} \mathcal{S}(\theta_s)
\end{aligned}
\tag{5}
$$

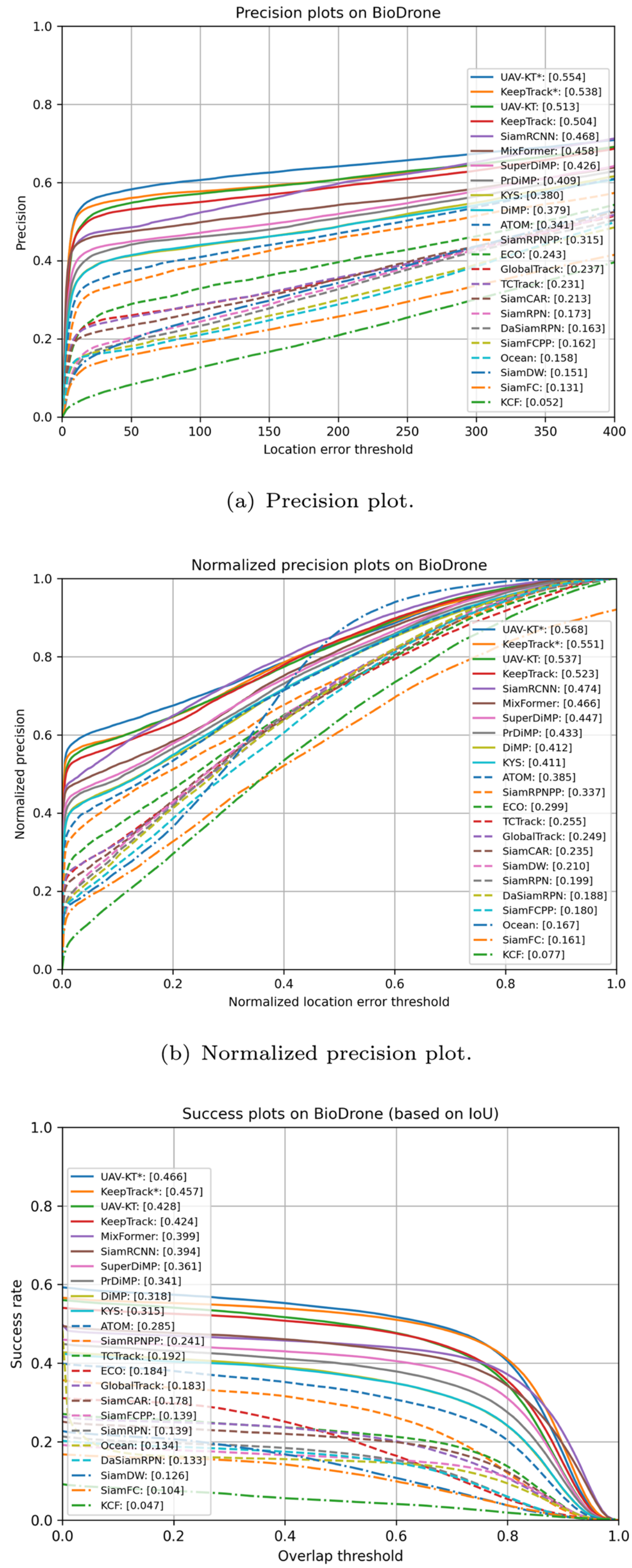
where $\Omega(\cdot)$ is the intersection over union. Recent work (Hu et al. 2023) also implements two more success scores based on generalized IoU (GIoU Rezatofighi et al. 2019) and distance IoU (DIoU Zheng et al. 2020) for calculation. Frames with overlap $s_t \geq \theta_s$ are defined as successful tracking. Draw the results based on various overlap threshold θ_s into a curve is the *success plot*, where the mAO (mean average overlap) is widelyused to rank trackers (S_{score}).

Robustness in R-OPE. The *robust plot* aims to exhibit the performance of trackers in the R-OPE mechanism. Each sequence is divided into several segments by the tracker's re-initialization points, thus the longest sub-sequence that a tracker successfully runs and the re-initialization points can be used to represent the robustness of the tracking process. Taking the number of restarts (R_{count}) and the average value of the longest sub-sequence L_{max} as abscissa and ordinate can generate a robust plot. Trackers closer to the upper left corner perform better (indicating successful tracking in longer sequences with rare re-initializations). Note that we do not limit the number of restarts under the R-OPE mechanism. Thus, we cannot only evaluate an algorithm by the above three metrics, since the high scores may be generated by frequent re-initializations. Therefore, the most reasonable metric for the R-OPE mechanism is the robustness plot and the number of restarts (R_{count}).

5.2 Performance of Generic SOT Trackers

We first compare the 20 represent trackers (Sect. 4.1) with the proposed baselines (Sect. 4.2) based on OPE and R-OPE evaluation mechanism, as shown in Table 3.

For OPE mechanism, precision plot, normalized precision plot, and success plot are selected for evaluation, as shown in Fig. 9. Except for the top-4 trackers which are all based on KeepTrack architecture (KeepTrack Mayer et al. 2021 and three proposed new baselines), we note that two



(a) Precision plot.

(b) Normalized precision plot.

(c) Success plot.

Fig. 9 General experiments of BioDrone based on OPE mechanism, evaluated by precision plot (**a**), normalized precision plot (**b**), and success plot (**c**). In brackets, we rank trackers by P_{score}, $P^{'}_{score}$, and S_{score}

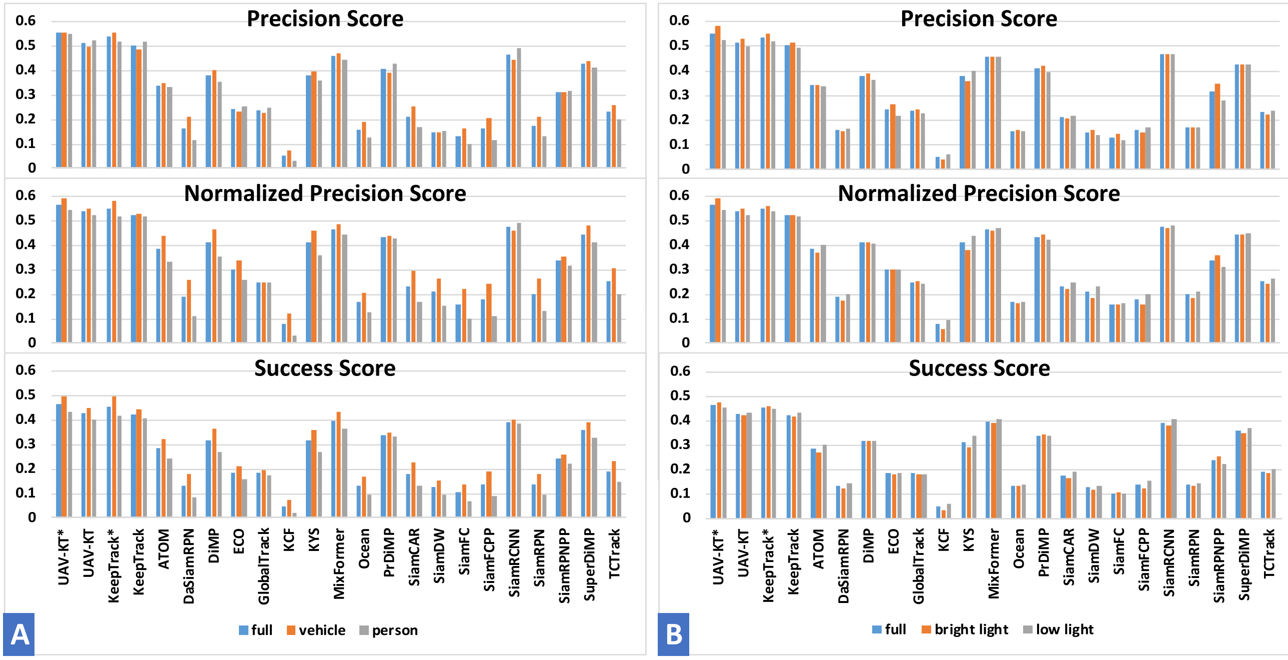


Fig. 10 General experiments of BioDrone based on OPE mechanism, evaluated in different target categories (**A**) and different light condition (**B**)

other trackers with different model architectures also perform well. MixFormer (Cui et al. 2022), a simple end-to-end model based on transformer structure, performs well in all evaluation metrics, indicating that the Mixed Attention Module (MAM) and a straightforward detection head can provide powerful tracking ability. Another re-detection-based model SiamRCNN (Voigtlaender et al. 2020) combines a two-stage scheme with a new trajectory-based dynamic planning algorithm and also achieves suitable tracking scores.

We also test trackers on two categories of targets (i.e., vehicles and persons) and three illumination conditions (i.e., bright light, low light (evening), and low light (night)). We combine low light (evening) and low light (night) into a single category and represented the test results in the above figure. In relation to different categories of moving targets (Fig. 10A), most algorithms exhibit better tracking performance on vehicles compared to persons. One possible explanation is that, from the perspective of a flapping-wing UAV, the size of a person is smaller than that of a vehicle, leading to a reduced number of available visual features and decreased robustness of the trackers. In various lighting conditions (Fig. 10B), most algorithms demonstrate superior tracking performance under bright light compared to low light. This indicates that inadequate lighting conditions diminish the visual features of moving targets and present challenges to the robustness of tracking.

Distinguished from the OPE mechanism, the R-OPE mechanism measures robust tracking capability mainly by the number of restarts. As shown in Fig. 11 and Table 3, all trackers perform better than the original OPE mechanism thanks to the re-initialization. However, all generic SOT trackers need more than 3 times re-initialization in tracking one BioDrone sequence, which means their robust tracking performances are limited in a very short period.

Moreover, we note that the series of methods based on combining CF and SNN (e.g., KeepTrack Mayer et al. 2021, SuperDiMP Danelljan et al. 2020, PrDiMP Danelljan et al. 2020, DiMP Bhat et al. 2019, ATOM Danelljan et al. 2019) are superior to the SNN-based algorithms (e.g., SiamRPN++ Li et al. 2019, SiamCAR Guo et al. 2020, SiamFC++ Xu et al. 2020, DaSiamRPN Zhu et al. 2018, SiamRPN Li et al. 2018, SiamDW Zhang and Peng 2019, SiamFC Bertinetto et al. 2016) of the same period in both OPE and R-OPE mechanisms. A possible reason is that most SNN-based methods exclude the update mechanism, and highly rely on the integrity of appearance and motion information. The tracking process is executed by matching features between the template region and the search region, while *tiny target* and *fast motion* can decrease the available target information, causing the SNN-based trackers to lose the target easily. On the contrary, the CF and SNN combination can take advantage of offline training and online updating, helping trackers to suit the appearance variations in the tracking process, and that is why we select the best CF-SNN combination tracker KeepTrack (Mayer et al. 2021) as our base model (Fig. 12).

5.3 Performance of the Proposed Baselines

Obviously, UAV-KT* and KeepTrack*, the two trackers which have been re-trained on the BioDrone benchmark,

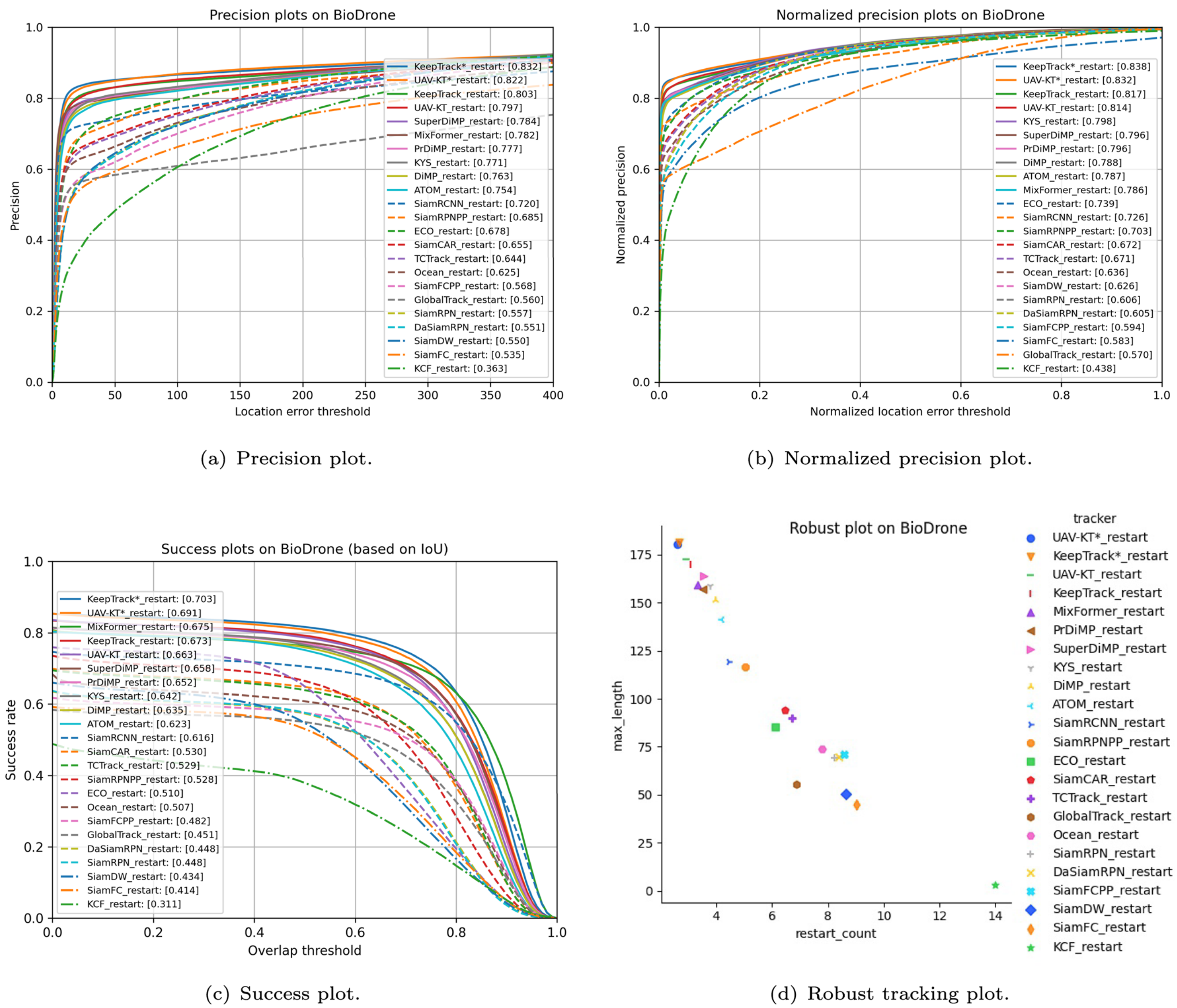


Fig. 11 General experiments of BioDrone based on the R-OPE mechanism, evaluated by precision plot (**a**), normalized precision plot (**b**), and success plot (**c**). In brackets, we rank trackers by P_{score}, $P^{'}_{score}$, and S_{score}. **d** Besides, BioDrone counts the number of restarts for each video, divides the entire video into several segments based on the restart point, and returns the longest sub-sequence that the algorithm successfully runs. Taking the number of restarts and the mean value of the longest sub-sequence as abscissa and ordinate can generate a robust plot. Trackers closer to the upper left corner perform better (indicating successful tracking in longer sequences with rare re-initializations)

achieve the best two performances in both OPE (Fig. 9) and R-OPE mechanisms (Fig. 11). For all trackers that have not been re-trained on BioDrone (we use the parameters and confirmations provided by the original authors), the proposed new baseline UAV-KT performs well. Here we design several ablation experiments to better exhibit the performance of the proposed new baseline UAV-KT and the training strategies.

5.3.1 Target Candidate Matching Network

The proposed UAV-KT utilizes some shallow features, which is especially effective for tiny targets, to obtain more meaningful features at the candidate embedding module. The score matrices are summed through the learned weights by the candidate matching module. Here, the weights are finally learned as [0.4929, 0.5070], in which the former is the shallow score

Fig. 12 Qualitative results of KeepTrack (Mayer et al. 2021) and the proposed baselines on BioDrone under the OPE mechanism (green bounding-box represents ground-truth, yellow bounding-box represents KeepTrack (Mayer et al. 2021), blue bounding-box represents UAV-KT, violet bounding-box represents KeepTrack*, red bounding-box represents UAV-KT*). Compared to the base model, UAV-KT* performs better when facing challenges in BioDrone (Color figure online)

matrix summing coefficient. Table 4(a) illustrates the performance of the original KeepTrack (Mayer et al. 2021) and the proposed UAV-KT. Note that neither of the two trackers is re-trained on BioDrone. Obviously, based on the target candidate matching network, UAV-KT improves its robustness by perceiving targets of different scales.

5.3.2 Different Training Strategies

As shown in Fig. 7, the original KeepTrack and UAV-KT include several parts (i.e., the base tracker, the target candidate extraction, and the target candidate association network). We notice that end-to-end training is not an appropriate strategy. Thus, to find a better training method, We design several strategies to explore the optimal parameters.

- *Strategy-1* re-train on the base tracker (KeepTrack*).
- *Strategy-2* Train target candidate association network with data from LaSOT and BioDrone training sets that meet the candidate conditions (KeepTrack#).

Table 4(b) shows that using BioDrone to re-train the base tracker improves the performance of KeepTrack (KeepTrack*), while the candidate association network performs poorly after re-training by the supplementary dataset (KeepTrack#).

We believe that this difference occurs because the two modules are designed for different tasks. (1) The task of the base tracker (SuperDiMP in KeepTrack) is target classification. A discriminative target predictor weight is obtained from template features, then it performs a cross-correlation operation with the frame features to be detected, and finally a score map is obtained. (2) Target candidate association network uses the score map from the base tracker to select target candidates, then extracts target candidate features for target candidate matching to finally identify the target.

Thus, using Strategy-1 for the base tracker can effectively improve the model's discriminative ability between forward and backward information, making it locate the target more robustly.

On the contrary, when Strategy-2 is applied to the target candidate association network, we first run the base tracker on all sequences of the BioDrone train-set to obtain tracking

Table 4 Ablation experiments of the proposed new baseline UAV-KT, based on the OPE mechanism

Tracker	P_{score} ↑	P'_{score} ↑	S_{score} ↑
(a) Performance of the new target candidate matching module			
KeepTrack (Mayer et al. 2021)	0.504	0.523	0.424
UAV-KT	0.513 (0.009 ↑)	0.537 (0.014 ↑)	0.428 (0.004 ↑)
(b) Performance of different training strategies			
KeepTrack (Mayer et al. 2021)	0.504	0.523	0.424
KeepTrack*	0.538 (0.034 ↑)	0.551 (0.028 ↑)	0.457 (0.033 ↑)
KeepTrack#	0.496 (0.008 ↓)	0.520 (0.003 ↓)	0.417 (0.007 ↓)
(c) Performance of the combination results			
Tracker	P_{score} ↑	P'_{score} ↑	S_{score} ↑
KeepTrack (Mayer et al. 2021)	0.504	0.523	0.424
KeepTrack*	0.538 (0.034 ↑)	0.551 (0.028 ↑)	0.457 (0.033 ↑)
UAV-KT*	0.554 (0.050 ↑)	0.568 (0.045 ↑)	0.466 (0.042 ↑)

results, and then set the train-set into two parts: a *train-train* and a *train-val* set. These datasets contain several tracking situations: (1) The correct candidate object is selected as the target. (2) It is no longer possible to track the target because the target classifier score of the corresponding candidate object is below a threshold. (3) Tracking fails, which includes the correct target existing but not selected or there is no correct target and none of the target candidates is selected. The task of the target candidate association module includes learning how to distinguish between the target with distractors, and how to remediate wrong results when the base tracker fails. However, due to the tiny target challenge, the appearance information on targets and distractors in BioDrone is not obvious. Thus, this training strategy may cause even a negative impact on trackers (please refer to the worse performance of KeepTrack# in Table 4(b)).

Based on the above analyses, Strategy-1 is selected as the final training strategy.

5.3.3 Results of Our New Baseline

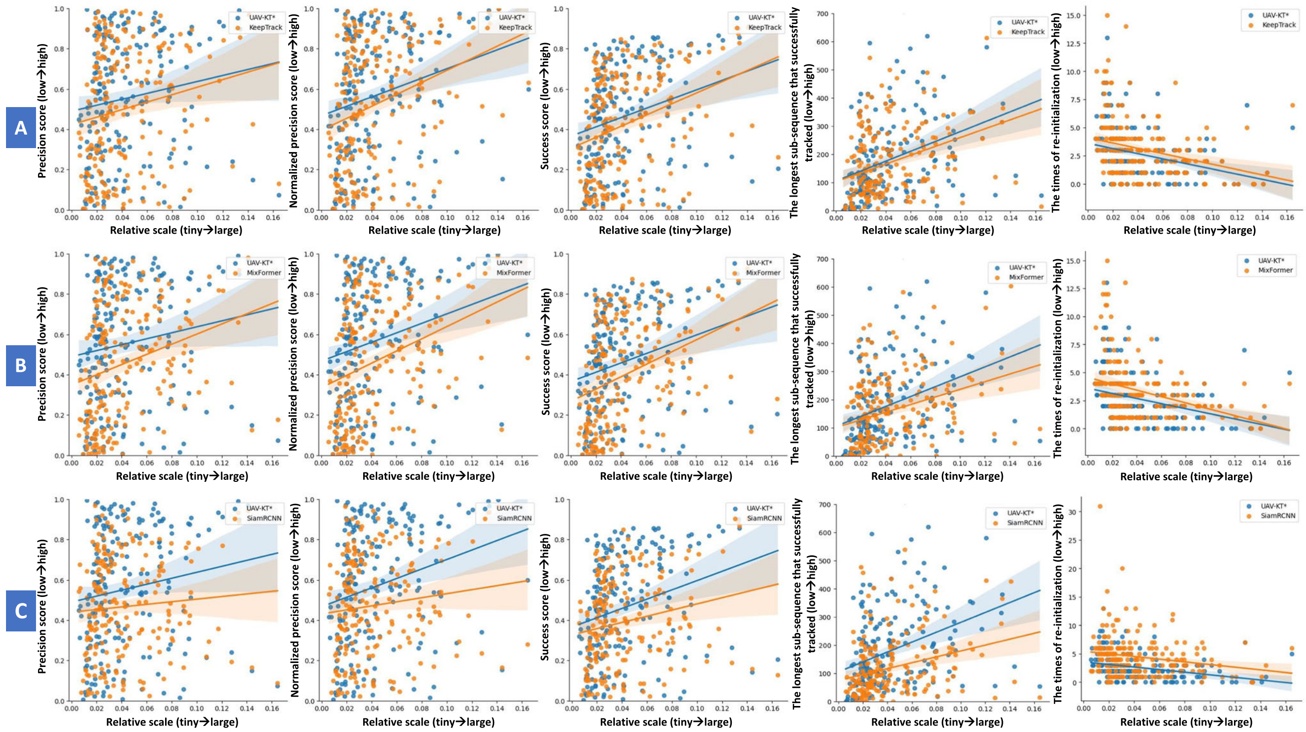
Our new baseline UAV-KT* employs the proposed target candidate association module and the training Strategy-1 based on the BioDrone. Table 4(c) illustrates that the combination improves the tracking performance effectively, which provides a novel direction for the following research.

5.4 Performance on Challenging Attributes

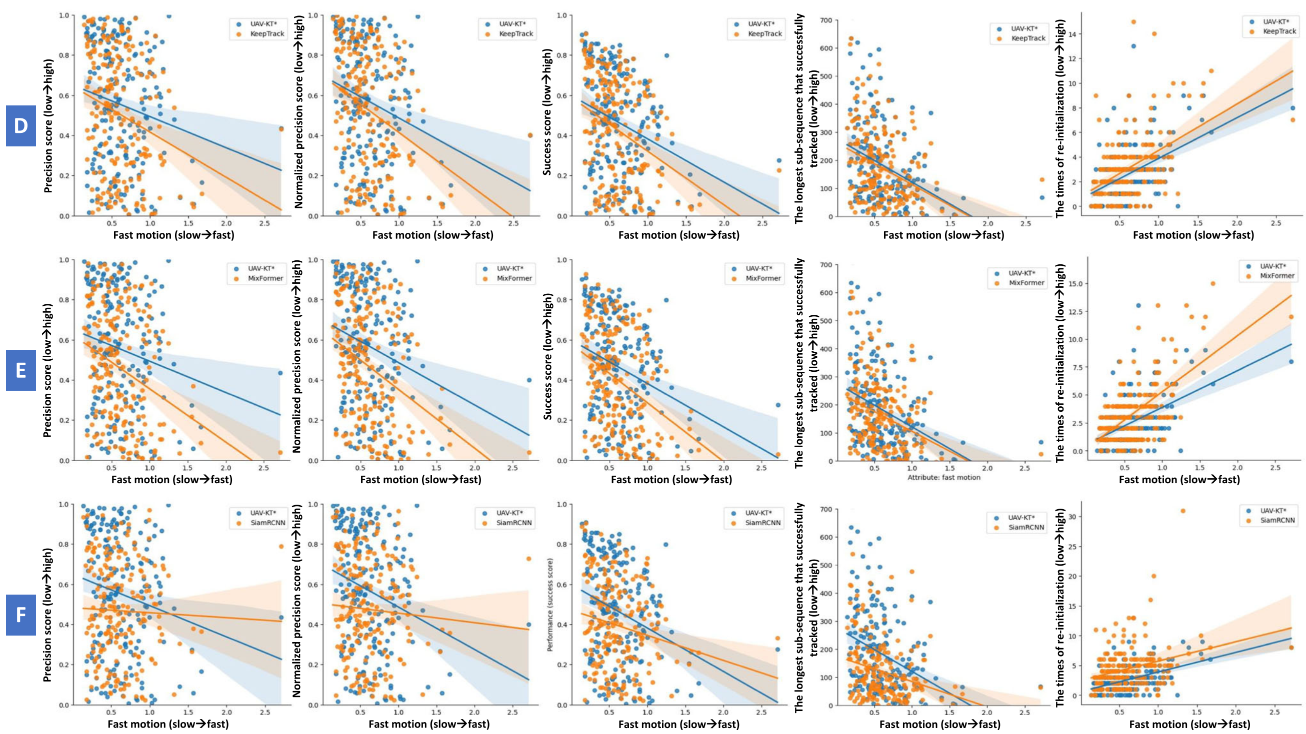
Different from tracking the target in generic scenarios, the UAV-based SOT task requires more visual robustness. In this section, we compare the proposed UAV-DT* baseline and three SOTA methods in challenging situations, to further analyze their robustness. Figure 13 illustrates the performance of trackers in tracking *tiny target* with *fast motion*. The above two factors reduce the available appearance information and abrupt the trajectories, causing trackers to fail easily.

Although SOTA methods like KeepTrack (Mayer et al. 2021), MixFormer (Cui et al. 2022), and SiamRCNN (Voigtlaender et al. 2020) perform well in generic situations (Fig. 1), they are easily failed in facing *tiny target*. Figure 13a shows that with the decrease in target size, performances of all trackers based on different mechanisms and metrics all drop quickly. For example, SiamRCNN (Voigtlaender et al. 2020) even fails more than 30 times in a sequence (the rightmost sub-figure in Fig. 13c), which shows that it is completely unable to handle this task, regardless of what strategies it has enabled. This phenomenon can also be observed in *fast motion* situation. As exhibited in Fig. 13b, the faster motion in two continuous frames, the poorer performance that trackers have.

Thus, the BioDrone benchmark introduces new challenging factors in the visual object tracking task and provides a comprehensive experimental environment for robust vision. Although existing methods perform poorly on this dataset, the proposed UAV-KT* gives a preliminary solution by optimizing the model structure and training strategies. However, some bad cases presented in Fig. 14 demonstrate that our base can be further improved, and multiple robust vision problems on BioDrone still deserve further research. The challenges brought by the *tiny target* and *fast motion* are highlighted in these examples. In contrast to tracking tasks in general scenes, pedestrians and vehicles appear significantly smaller in the drone's field of view. Additionally, the shaking and rotation of the camera during flapping flight can disturb the motion trajectory of the target, thus presenting significant challenges for algorithms that depend on visual features and motion information.



(a) Performance in tracking tiny target (smaller value in horizontal coordinate means including more tiny targets).



(b) Performance in tracking fast motion target (larger value in horizontal coordinate means including faster motion).

Fig. 13 Performance of the proposed UAV-KT* and represent generic SOT methods on challenging attributes. The scores of each algorithm in the test set (200 videos) are plotted as scatter plots. Where the vertical coordinates represent the scores of the algorithms (from left to right: precision score P_{score}, normalized precision score $P^{'}_{score}$, and success score S_{score} in OPE mechanism; the average value of the longest sub-sequence L_{max} and the number of restarts R_{count} in R-OPE mechanism). The horizontal coordinates of **a** represent the average relative target scale, and the horizontal coordinates of **b** represent the average target motion in a video. Clearly, UAV-KT* performs better than KeepTrack (Mayer et al. 2021), MixFormer (Cui et al. 2022), and SiamRCNN (Voigtlaender et al. 2020) in both tiny target and fast motion challenges

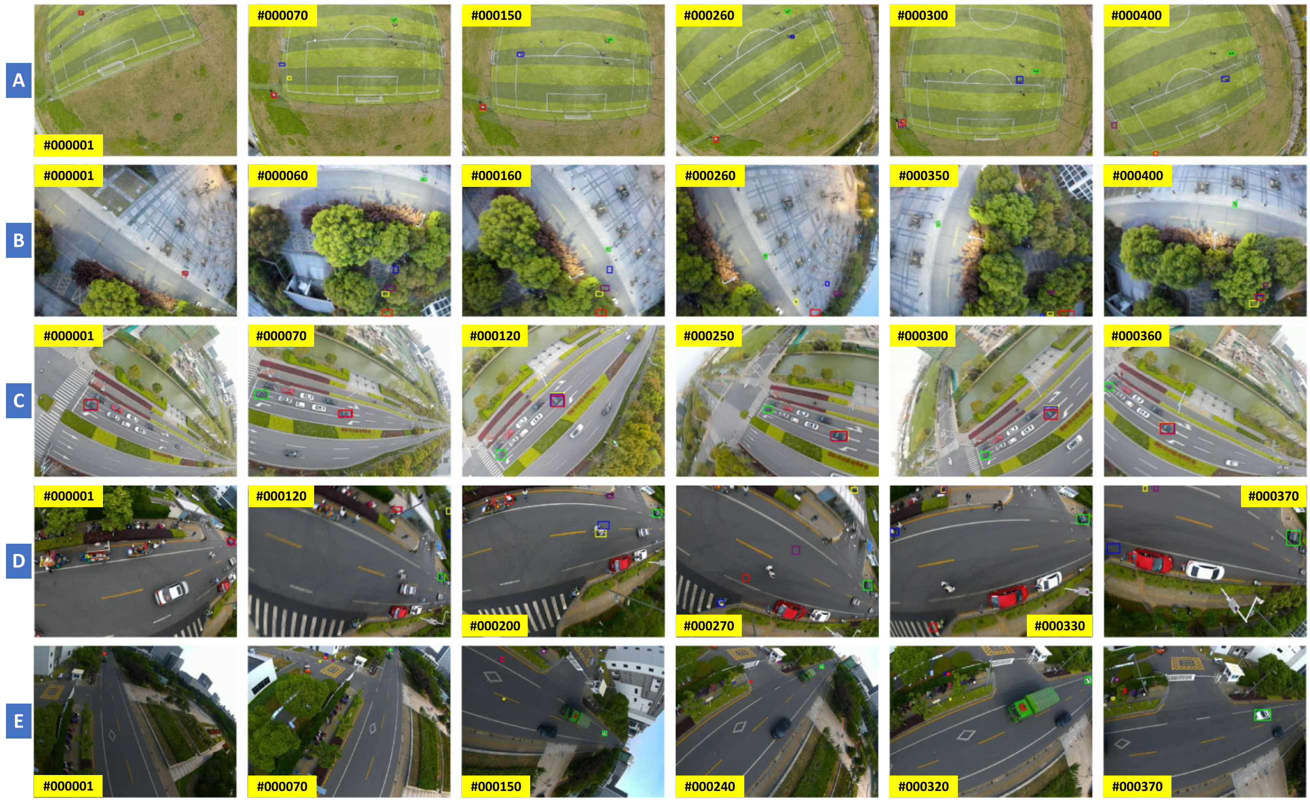


Fig. 14 Qualitative results of some bad cases for the represent trackers on OPE mechanism (■ green bounding-box represents ground-truth, ■ yellow bounding-box represents KeepTrack (Mayer et al. 2021), ■ blue bounding-box represents UAV-KT, ■ violet bounding-box represents KeepTrack*, ■ red bounding-box represents UAV-KT*) (Color figure online)

6 Conclusion

In this paper, a bionic drone-based single object tracking benchmark BioDrone is proposed for robust vision research. Unlike existing benchmarks that are mainly based on fixed-wing or rotary-wing UAVs, the flapping-wing system selected by BioDrone includes additional visual challenges due to its serious camera shake. Compared with existing works, BioDrone is the largest UAV-based SOT benchmark with a *smaller target size* and *more drastic appearance changes* between consecutive frames. It includes *600* videos with *304,209* manually labeled frames, and automatically generates frame-level labels for ten challenge attributes, which provides a high-quality and challenging experimental environment for robust vision research. Besides, We further optimize the SOTA method KeepTrack (Mayer et al. 2021) and design a new baseline UAV-KT with a suitable training strategy, aiming to propose a preliminary baseline for challenging factors in BioDrone. Finally, we test our method and 20 representative methods by comprehensive evaluation mechanisms and metrics in BioDrone, and experimental results indicate that the proposed method achieves 5% performance boost in the precision score. However, several failure cases and systematic analyses indicate that BioDrone still contains many unresolved challenges and deserves further attention in robust vision research.

In the future, we believe that the proposed BioDrone benchmark can provide a high-quality experimental environment for further research, and help researchers to design new robust tracking methods. Besides, this work also represents a broader range of SOT problems, such as those in high-speed autonomous driving, and egocentric vision. While BioDrone mainly focuses on bionic UAVs, the results and findings in this paper might transfer to those more comprehensive problems.

Availability of data and materials. All data will be made available on reasonable request.

Declarations

Conflict of interest All authors declare no conflicts of interest.

Code availability The toolkit and experimental results will be made publicly available.

References

Abu Alhaija, H., Mustikovela, S. K., Mescheder, L., Geiger, A., & Rother, C. (2018). Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision, 126*(9), 961–972.

Barrientos, A., Colorado, J., Martinez, A., & Valente, J. (2010). Rotary-wing mav modeling and control for indoor scenarios. In *2010 IEEE international conference on industrial technology* (pp. 1475–1480). IEEE.

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *European conference on computer vision* (pp. 850–865). Springer.

Bhat, G., Danelljan, M., Gool, L. V., & Timofte, R. (2019). Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6182–6191).

Bhat, G., Danelljan, M., Gool, L. V., & Timofte, R. (2020). Know your surroundings: Exploiting scene information for object tracking. In *European conference on computer vision* (pp. 205–221). Springer.

Bondi, E., Dey, D., Kapoor, A., Piavis, J., Shah, S., Fang, F., Dilkina, B., Hannaford, R., Iyer, A., Joppa, L., et al. (2018). Airsim-w: A simulation environment for wildlife conservation with uavs. In *Proceedings of the 1st ACM SIGCAS conference on computing and sustainable societies* (pp. 1–12).

Bondi, E., Jain, R., Aggrawal, P., Anand, S., Hannaford, R., Kapoor, A., Piavis, J., Shah, S., Joppa, L., & Dilkina, B., et al. (2020). Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision* (pp. 1747–1756).

Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., & Fu, C. (2022). Tctrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14798–14808).

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531.

Cui, Y., Jiang, C., Wang, L., & Wu, G. (2022). Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13608–13618).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886–893). IEEE.

Danelljan, M., Bhat, G., Khan, F. S., & Felsberg, M. (2019). Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4660–4669).

Danelljan, M., Bhat, G., Shahbaz Khan, F., & Felsberg, M. (2017). Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6638–6646).

Danelljan, M., Gool, L. V., & Timofte, R. (2020). Probabilistic regression for visual tracking. In *2020 IEEE conference on computer vision and pattern recognition (CVPR)*.

De Croon, G., Perçin, M., Remes, B., Ruijsink, R., & De Wagter, C. (2016). *The delfly* (pp. 978–94). Dordrecht: Springer.

Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., & Leal-Taixé, L. (2021). Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision, 129*(4), 845–881.

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 224–236).

Dupeyroux, J., Serres, J. R., & Viollet, S. (2019). Antbot: A six-legged walking robot able to home like desert ants in outdoor environments. *Science Robotics, 4*(27), eaau0307.

Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Huang, M., Liu, J., Xu, Y., et al. (2021). Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision, 129*(2), 439–461.

Finlayson, G. D., & Trezzi, E. (2004). Shades of gray and colour constancy. In *The twelfth color imaging conference 2004* (pp. 37–41).

Fraire, A. E., Morado, R. P., López, A. D., & Leal, R. L. (2015). Design and implementation of fixed-wing mav controllers. In *2015 Workshop on research, education and development of unmanned aerial systems (RED-UAS)* (pp. 172–179). IEEE.

Gauglitz, S., Höllerer, T., & Turk, M. (2011). Evaluation of interest point detectors and feature descriptors for visual tracking. *International Journal of Computer Vision, 94*(3), 335–360.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

Guo, D., Wang, J., Cui, Y., Wang, Z., & Chen, S. (2020). Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6269–6277).

Han, L., Wang, P., Yin, Z., Wang, F., & Li, H. (2021). Context and structure mining network for video object detection. *International Journal of Computer Vision, 129*(10), 2927–2946.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2014). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(3), 583–596.

Hsieh, M.-R., Lin, Y.-L., & Hsu, W. H. (2017). Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision* (pp. 4145–4153).

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., & Markham, A. (2022). Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision, 130*(2), 316–343.

Hu, S., Zhao, X. & Huang, K. (2023). SOTVerse: A user-defined task space of single object tracking. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-023-01908-5.

Hu, S., Zhao, X., Huang, L., & Huang, K. (2023). Global instance tracking: Locating target more like humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(1), 576–592.

Huang, L., Zhao, X., & Huang, K. (2020). Globaltrack: A simple and strong baseline for long-term tracking. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 11037–11044).

Huang, L., Zhao, X., & Huang, K. (2021). Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(5), 1562–1577.

Jiang, B., Luo, R., Mao, J., Xiao, T., & Jiang, Y. (2018). Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 784–799).

Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision, 130*(5), 1366–1401.

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin, L., Vojír, T., Häger, G., Lukežič, A., Fernández, G., Gupta, A., Petrosino, A., Memarmoghadam, A., Garcia-Martin, A., Solís Montero, A., et al. (2016). The visual object tracking

VOT2016 challenge results. In *Computer vision—ECCV 2016 workshops* (pp. 777–823). Springer.

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.-K., Danelljan, M., Zajc, L.Č., Lukežič, A., Drbohlav, O., He, L., et al. (2020). The eighth visual object tracking VOT2020 challenge results. In *Computer vision—ECCV 2020 workshops* (pp. 547–601). Springer.

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Zajc, L. Č., Vojír, T., Bhat, G., Lukežič, A., Eldesokey, A., Fernández, G., et al. (2019a). The sixth visual object tracking VOT2018 challenge results. In *Computer vision—ECCV 2018 workshops* (pp. 3–53). Springer.

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Zajc, L. C., Vojír, T., Häger, G., Lukežic, A., Eldesokey, A., Fernández, G., García-Martín, Á., Muhic, A., Petrosino, A., Memarmoghadam, A., et al. (2017). The visual object tracking VOT2017 challenge results. In *Proceedings of 2017 IEEE international conference on computer vision workshops (ICCVW)* (pp. 1949–1972). IEEE, Venice, Italy.

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G., Vojir, T., Hager, G., Nebehay, G., Pflugfelder, R., Gupta, A., Bibi, A., Lukezic, A., Garcia-Martin, A., Saffari, A., Petrosino, A., & Solis Montero, A. (2015). The visual object tracking VOT2015 challenge results. In *Proceedings of 2015 IEEE international conference on computer vision workshop (ICCVW)* (pp. 564–586). IEEE.

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kämäräinen, J.-K., Cehovin Zajc, L., Drbohlav, O., Lukezic, A., Berg, A., Eldesokey, A., Käpylä, J., Fernández, G., Gonzalez-Garcia, A., Memarmoghadam, A., et al. (2019b). The seventh visual object tracking VOT2019 challenge results. In *Proceedings of 2019 IEEE/CVF international conference on computer vision workshop (ICCVW)* (pp. 2206–2241). IEEE, Seoul, Korea (South).

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kämäräinen, J.-K., Chang, H. J., Danelljan, M., Zajc, L. Č., Lukežič, A., Drbohlav, O., et al. (2021). The ninth visual object tracking VOT2021 challenge results. In *Proceedings of 2021 IEEE/CVF international conference on computer vision workshops (ICCVW)* (pp. 2711–2738). IEEE, Montreal, BC, Canada.

Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Cehovin, L., Nebehay, G., Fernandez, G., Vojir, T., Gatt, A., Khajenezhad, A., Salahledin, A., Soltani-Farani, A., et al. (2013). The visual object tracking VOT2013 challenge results. In *Proceedings of 2013 IEEE international conference on computer vision workshops (ICCVW)* (pp. 98–111). IEEE.

Kristan, M., Pflugfelder, R. P., Leonardis, A., Matas, J., Cehovin, L., Nebehay, G., Vojír, T., Fernández, G., Lukezic, A., Dimitriev, A., Petrosino, A., Saffari, A. A., et al. (2014). The visual object tracking VOT2014 challenge results. In L. Agapito, M. M. Bronstein, & C. Rother (Eds.), *Computer vision: ECCV 2014 workshops* (Vol. 8926, pp. 191–217). Springer.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* 25.

Lee, N., Lee, S., Cho, H., & Shin, S. (2018). Effect of flexibility on flapping wing characteristics in hover and forward flight. *Computers & Fluids, 173*, 111–117.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., & Yan, J. (2019). Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4282–4291).

Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with siamese region proposal network. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Li, S., & Yeung, D.-Y. (2017). Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Thirty-first AAAI conference on artificial intelligence*.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International Journal of Computer Vision, 128*(2), 261–318.

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., & Leibe, B. (2021). Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision, 129*(2), 548–578.

Mayer, C., Danelljan, M., Paudel, D.P., & Van Gool, L. (2021). Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13444–13454).

McMasters, J., & Cummings, R. (2004). Rethinking the airplane design process: An early 21st century perspective. In *42nd AIAA aerospace sciences meeting and exhibit* (p. 693).

McMasters, J. H., & Cummings, R. M. (2002). Airplane design: Past, present, and future. *Journal of Aircraft, 39*(1), 10–17.

Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., & Ghanem, B. (2018). Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 300–317).

Müller, M., Casser, V., Lahoud, J., Smith, N., & Ghanem, B. (2018). Sim4cv: A photo-realistic simulator for computer vision applications. *International Journal of Computer Vision, 126*(9), 902–919.

Mueller, M., Smith, N., & Ghanem, B. (2016). A benchmark and simulator for uav tracking. In *European conference on computer vision* (pp. 445–461). Springer.

Pech-Pacheco, J. L., Cristobal, G., Chamorro-Martinez, J., & Fernandez-Valdivia, J. (2000). Diatom autofocusing in brightfield microscopy: A comparative study. In *Proceedings 15th international conference on pattern recognition. ICPR-2000* (Vol. 3, pp. 314–317).

Pornsin-Sirirak, T. N., Tai, Y.-C., Ho, C.-M., & Keennon, M. (2001). Microbat: A palm-sized electrically powered ornithopter. In *Proceedings of NASA/JPL workshop on biomorphic robotics* (Vol. 14, p. 17). Citeseer.

Ramakrishnan, S. K., Jayaraman, D., & Grauman, K. (2021). An exploration of embodied visual exploration. *International Journal of Computer Vision, 129*(5), 1616–1649.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 28.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 658–666).

Rigelsford, J. (2004). Neurotechnology for biomimetic robots. *Industrial Robot: An International Journal, 31*(6), 534.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Ryu, S., Kwon, U., & Kim, H. J. (2016). Autonomous flight and vision-based target tracking for a flapping-wing mav. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 5645–5650). IEEE.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4938–4947).

Sims, C. A., & Uhlig, H. (1991). Understanding unit rooters: A helicopter tour. *Econometrica: Journal of the Econometric Society, 59*, 1591–1599.

Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781–10790).

Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627–9636).

Van De Weijer, J., Schmid, C., Verbeek, J., & Larlus, D. (2009). Learning color names for real-world applications. *IEEE Transactions on Image Processing, 18*(7), 1512–1523.

Voigtlaender, P., Luiten, J., Torr, P. H., & Leibe, B. (2020). Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6578–6588).

Wu, X., Li, W., Hong, D., Tao, R., & Du, Q. (2021). Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine, 10*(1), 91–124.

Wu, Y., Lim, J., & Yang, M.-H. (2013). Online object tracking: A benchmark. In *2013 IEEE conference on computer vision and pattern recognition* (pp. 2411–2418).

Wu, Y., Lim, J., & Yang, M.-H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1834–1848.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2018). Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3974–3983).

Xu, Y., Wang, Z., Li, Z., Yuan, Y., & Yu, G. (2020). Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 12549–12556).

Yang, W., Wang, L., & Song, B. (2018). Dove: A biomimetic flapping-wing micro air vehicle. *International Journal of Micro Air Vehicles, 10*(1), 70–84.

Yu, H., Li, G., Zhang, W., Huang, Q., Du, D., Tian, Q., & Sebe, N. (2020). The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *International Journal of Computer Vision, 128*(5), 1141–1159.

Zhang, C., & Rossi, C. (2017). A review of compliant transmission mechanisms for bio-inspired flapping-wing micro air vehicles. *Bioinspiration & Biomimetics, 12*(2), 025005.

Zhang, Z., & Peng, H. (2019). Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4591–4600).

Zhang, Z., Peng, H., Fu, J., Li, B., & Hu, W. (2020). Ocean: Object-aware anchor-free tracking. In *European conference on computer vision* (pp. 771–787). Springer.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 12993–13000).

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2021). Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 7380–7399.

Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., & Hu, W. (2018). Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 101–117).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.