Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080050

REVIEW

A Challenge-Driven Survey on UAV-Based Target Tracking

Lingyu Jin[1,2], **Rui Wang**[1,2] **and Bo Huang**[1,2,*]

[1]College of Optoelectronic Engineering, Chongqing University, Chongqing, China
[2]Key Laboratory of Optoelectronic Technology and Systems of the Education Ministry of China, Chongqing University, Chongqing, China

*Corresponding Author: Bo Huang. Email: huangbo0326@cqu.edu.cn

Received: 02 February 2026; Accepted: 23 March 2026; Published: 08 May 2026

ABSTRACT: Unmanned Aerial Vehicle (UAV) target tracking is one of the key technologies in aerial intelligent perception systems, playing a vital role in applications such as traffic monitoring, border patrol, disaster response, search and rescue, environmental monitoring, and military reconnaissance. Compared with generic object tracking tasks, UAV platforms exhibit significant differences in imaging perspectives, target scales, motion patterns, and onboard computing capabilities, which pose unique challenges for UAV target tracking, including small targets and drastic scale variations, platform motion and motion blur, complex backgrounds and frequent occlusions, low-light conditions at night, as well as real-time and energy constraints. To address these issues, a large number of UAV-oriented tracking methods have been proposed in recent years, covering traditional correlation filters, deep Siamese networks, and emerging Transformer-based models, achieving continuous performance improvements across multiple UAV benchmark datasets. Despite substantial research efforts, existing survey works primarily focus on generic object tracking or a single technical approach, lacking a systematic summary and comparative analysis from the perspective of UAV application requirements. Unlike previous surveys that mainly classify methods based on model architecture, the innovation of this study lies in establishing a unified UAV target tracking framework centered on five major challenges. First, we analyze typical UAV tracking applications and core challenges. Then, from a challenge-driven perspective, existing methods are categorized and summarized based on small targets and scale variations, rapid motion and motion blur, complex backgrounds and occlusions, low-light night conditions, and lightweight and real-time considerations. Furthermore, we conduct quantitative and qualitative comparisons of representative methods in terms of accuracy, success rate, and computational efficiency on mainstream benchmarks, including UAV123, UAV123@10fps, UAVDT, UAV20L, and DTB70. Finally, we looked ahead to future development directions, such as lightweight deployment, multi-modal fusion, and large model-driven approaches. This work aims to provide a clear technical roadmap and a systematic reference for UAV target tracking research.

KEYWORDS: UAV target tracking; correlation filter (CF); siamese network (SN); transformer; comprehensive survey

1 Introduction

Visual object tracking is a fundamental research problem in the field of computer vision and plays a crucial role in a wide range of applications, including video surveillance [1], augmented reality [2], robotics [3], autonomous navigation [4], and marine exploration [5]. In recent years, with the rapid development of Unmanned Aerial Vehicle (UAV) technology, visual object tracking has gradually become a core component of aerial vision systems, enabling UAVs to continuously perceive, localize, and follow targets in complex dynamic environments [6,7].

Compared with traditional ground-based vision systems, UAV platforms possess advantages such as high mobility, flexible deployment, and wide-area coverage, and have been extensively applied in traffic monitoring, border patrol, disaster response, search and rescue missions, environmental monitoring, and military reconnaissance [8–10]. In these applications, visual object tracking serves as a key perception module that allows UAVs to continuously localize targets over time, thereby providing essential support for autonomous navigation, path planning, and active control [6]. However, due to differences in imaging viewpoints and platform characteristics, UAV target tracking faces challenges that are far more severe than those encountered in conventional generic object tracking tasks [11].

Fig. 1 illustrates the distinctive characteristics and challenges of UAV target tracking in comparison with generic object tracking based on fixed or ground-mounted cameras.

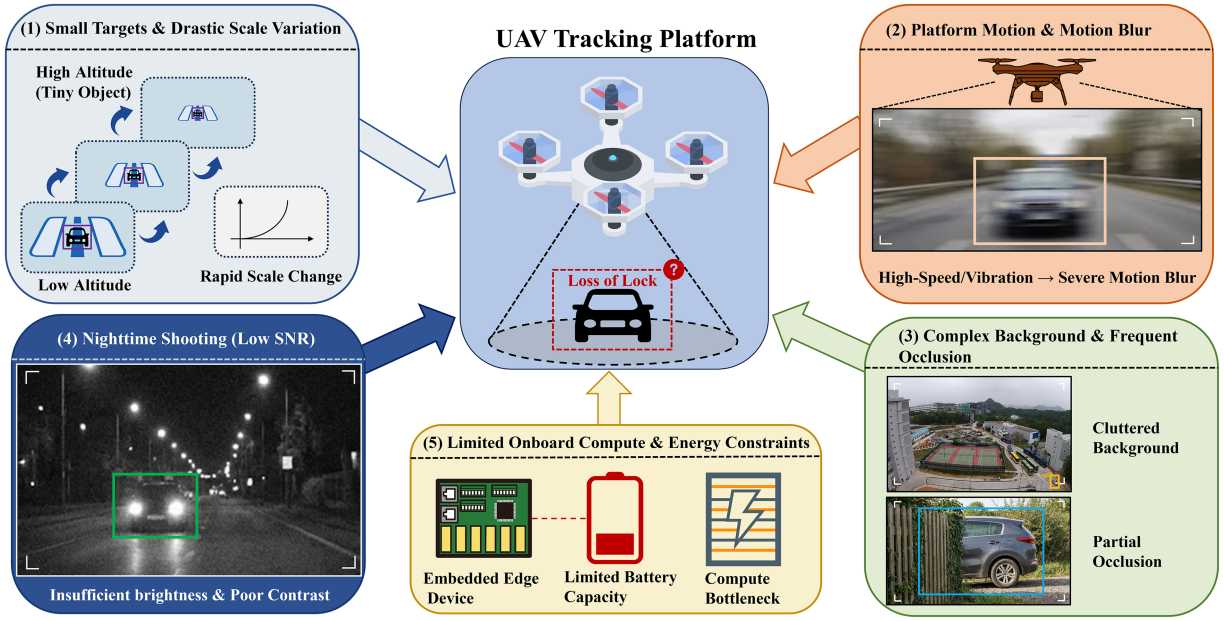

Figure 1: Key challenges imposed on UAV-based visual object tracking systems.

(1) **Small target size and drastic scale variations:** Owing to changes in flight altitude and long-distance imaging, targets in UAV videos are typically small in size and undergo frequent scale variations, which degrade feature representation capability and increase the risk of tracking failure [12,13]. Compared with generic tracking where the camera is often static or slow-moving, UAVs usually capture a much larger field of view, introducing more background content and effectively reducing the visual resolution of the target. As a result, fine-grained cues that are reliable in generic settings become weak or even ambiguous in aerial views, making scale estimation and identity discrimination substantially more difficult under the same model capacity.

(2) **Platform motion and motion blur:** The high-speed motion, attitude changes, and environmental disturbances of UAVs can cause abrupt viewpoint changes and motion blur, making the appearance and motion patterns of targets difficult to model [13]. In UAV tracking, both the platform and the target may move rapidly, and mechanical vibration or strong wind can further aggravate blur, which is less common in many ground-camera benchmarks. This combination enlarges inter-frame displacement and increases the probability that the target leaves the search region, thereby amplifying drift for trackers that rely on local matching.

(3) **Complex backgrounds and frequent occlusions:** Aerial videos usually contain complex backgrounds with numerous similar distractors, while occlusions caused by buildings, trees, and other obstacles occur frequently [14]. Because UAV imagery often covers wide areas, background clutter and look-alike objects are more prevalent, which weakens the discriminative margin between the target and its surroundings. Moreover, occlusion can be partial or even full, directly corrupting the target template and making recovery difficult without explicit re-detection or global reasoning mechanisms.

(4) **Night-time scenarios:**Under night-time or low-light conditions, images captured by UAVs often suffer from insufficient illumination, reduced contrast, and loss of effective texture information, leading to degraded target appearance features and a significant reduction in tracking robustness [15]. In practice, illumination can change rapidly from bright to dim due to shadows, canopy regions, or sudden exposure shifts during flight, causing pronounced appearance inconsistency across frames. When low light co-occurs with small targets and motion blur, the target becomes less distinguishable from the background, and conventional RGB trackers are more likely to fail without dedicated enhancement or robust representation learning.

(5) **Limited onboard computational resources and energy constraints:** Limited onboard computational resources and energy constraints: UAVs are restricted by payload capacity and battery life, making it difficult to deploy computationally intensive models. Therefore, tracking algorithms must strike a balance between accuracy and real-time performance on embedded UAV platforms [13,15]. This constraint makes the evaluation and design criteria different from generic tracking, where high-end GPUs are commonly used and efficiency is often secondary. For UAV deployment, achieving robust tracking under the above challenges is important, but finding a practical trade-off between performance and efficiency is even more critical to ensure stable onboard operation.

These challenges indicate that UAV target tracking cannot be simply regarded as an extension of generic object tracking, but instead requires dedicated designs that take the characteristics of UAV platforms into account. To address the aforementioned challenges, a variety of UAV-oriented target tracking methods have been proposed. Early studies on object tracking mainly relied on traditional approaches, including Kalman filtering [16], MeanShift tracking [17], and correlation filter-based methods [18]. Kalman filtering [16] predicts the target state with a linear motion model and updates it using noisy observations, which is efficient and stable under smooth motion, but it can be inaccurate when UAV motion induces abrupt maneuvers, strong nonlinear dynamics, or intermittent observations. MeanShift tracking [17] performs local mode seeking on an appearance density, which is lightweight and effective for gradual appearance variation, yet it is sensitive to background clutter because similar distractors can dominate the local density peak in aerial views. Correlation filter-based methods [8] learn a discriminative template efficiently and can run at very high speed, which is attractive for onboard deployment, but they are prone to drift under long occlusion and complex background interference because local response peaks may be unreliable and online updates may incorporate wrong target information.

With the rapid development of deep learning, visual object tracking methods have achieved significant progress in feature representation capability and adaptability to complex environments. Representative tracking algorithms based on deep Siamese networks and discriminative learning, such as SiamFC [19], SiamRPN [20], SiamR-CNN [21], SiamAPN [22], SiamAPN++ [23], and SiamDT [14], effectively improve tracking performance in complex backgrounds and scenarios with scale variations through end-to-end similarity matching or discriminative modeling mechanisms. SiamFC [19] formulates tracking as template-to-search similarity matching, which enables efficient end-to-end training and strong generalization, but in UAV videos it can lose targets under large inter-frame displacement because the target may move outside the search region. SiamRPN [20] integrates region proposal and box regression into the Siamese pipeline, which improves localization precision and scale handling, yet it still relies on the assumption that the target remains

within a bounded search area. SiamR-CNN [21] introduces stronger discrimination through proposal-based refinement, which helps suppress distractors in cluttered aerial scenes and improves robustness under partial occlusion, but its refinement stages usually increase computation and latency. SiamAPN [22] enhances UAV-oriented localization by strengthening alignment between features and target geometry, which is beneficial for small targets and rapid scale changes, but its performance may still degrade when motion blur and severe viewpoint changes coexist. SiamAPN++ [23] further refines the UAV-oriented design by improving matching and localization stability, which helps reduce drifting in complex backgrounds, yet it remains sensitive when the target is both tiny and heavily occluded. SiamDT [14] explicitly targets anti-distractor tracking by strengthening discrimination against similar objects, which is crucial for aerial scenes with repetitive textures and look-alike objects, but maintaining strong discrimination often increases model complexity and can challenge real-time deployment.

In recent years, Transformer-based tracking methods, including TCTrack [24], HiFT [25], SGDViT [26], STARK [27], TrDiMP [28], MixFormer [29], MixFormerV2 [30], and SwinTrack [31], have further introduced global feature modeling capabilities, demonstrating stronger potential in addressing UAV-specific challenges such as drastic appearance variations, occlusions, and complex motion patterns. TCTrack [24] improves robustness by enhancing information interaction across frames, which is helpful when UAV motion causes rapid appearance changes, but the benefit depends on stable temporal cues that may be weakened by heavy blur. HiFT [25] strengthens multi-scale feature interaction, which improves representation for small targets and scale variation, yet multi-scale interaction can increase computation when deployed onboard. SGDViT [26] focuses on improving attention to salient target cues under cluttered backgrounds, which is valuable for aerial scenes, but its effectiveness can be limited when the target occupies only a few pixels. STARK [27] introduces transformer-based feature interaction for stronger global reasoning, which helps suppress background interference and handle partial occlusion, but the cost of global modeling raises deployment difficulty on edge devices. TrDiMP [28] enhances discriminative learning with transformer-style feature interaction, which can improve robustness under appearance variation, yet it often requires careful efficiency design to satisfy real-time constraints. MixFormer [29] improves template-search interaction by mixing information more effectively, which stabilizes matching under viewpoint change and motion blur, but global interaction may still be expensive on UAV hardware. MixFormerV2 [30] further strengthens interaction and efficiency balance, which benefits challenging UAV scenarios, although the overall model complexity can remain high. SwinTrack [31] leverages hierarchical modeling to capture multi-scale structures, which is especially relevant for small targets and strong scale changes, but maintaining both accuracy and speed on embedded platforms remains a practical challenge. To reflect the most recent progress, several UAV-oriented trackers proposed in 2025 further strengthen both robustness and deployment feasibility. ORTrack [32] focuses on the frequent occlusion issue in aerial scenes by improving occlusion robustness in Transformer-based tracking, and its design aims to reduce the typical failure of drifting after partial occlusion. SSTrack [33] reports strong overall performance across multiple UAV benchmarks and represents the recent trend of improving representation and matching stability under complex backgrounds and rapid appearance variations. In addition, FWRDCF [34] shows that correlation-filter-based tracking remains an active direction in UAV scenarios, where carefully designed feature weighting and robust learning strategies can improve accuracy while preserving high efficiency.

However, deep models usually incur high computational costs, which pose stricter requirements for deployment on UAV platforms. In addition, multi-modal fusion-based tracking [35] has emerged to exploit complementary sensing cues for low-light and degraded conditions, but its robustness relies on sensor alignment and synchronization that are not always stable on UAVs. Lightweight model design [36] aims to reduce latency and resource consumption through compact backbones and efficient modules, which is

critical for onboard deployment, but aggressive compression can reduce representation capacity and hurt robustness under compound UAV challenges.

Despite substantial research efforts, existing survey works mainly focus on generic object tracking or a single category of methods, and lack a systematic summary tailored to UAV platforms. In particular, comprehensive analyses of emerging research directions remain insufficient. This paper aims to provide a systematic and comprehensive review of recent advances in UAV visual object tracking. Fig. 2 presents the logical relationships and technical contributions of the survey. The main contributions of this survey are summarized as follows:

- Systematically categorizes and summarizes UAV target tracking methods based on the specific challenges they address.
- Addresses the limitations of existing literature and fills the research gaps regarding emerging directions.
- Conducts a comprehensive analysis of key metrics, including precision, success rate, and frame rate, across various UAV tracking benchmarks.
- Outlines future development directions for UAV target tracking.

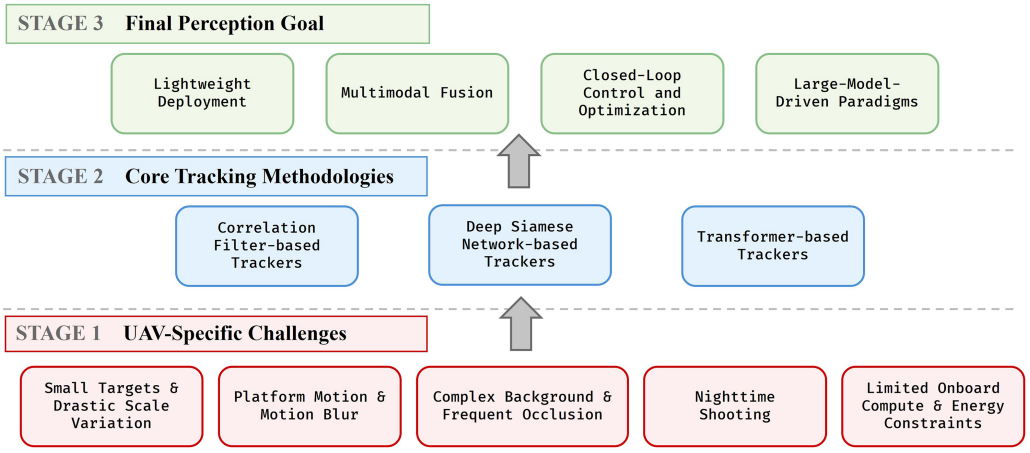

Figure 2: Workflow and contribution map of this survey.

2 Related Work

This survey follows a structured literature collection and screening procedure to ensure coverage and relevance for UAV-based visual target tracking. We conducted keyword-based searches in widely used academic databases and digital libraries (e.g., IEEE Xplore, ACM Digital Library, Web of Science/Scopus, and Google Scholar) and complemented them with arXiv for the most recent preprints. The search terms combined "UAV" (or "drone"/"aerial") with "visual tracking", "object tracking", "single-object tracking", "Siamese tracking", "correlation filter tracking", and "Transformer tracking", together with dataset keywords such as "UAV123", "UAVDT", and "DTB70". The primary time span considered is from 2014 to 2025, covering the evolution from correlation-filter trackers to modern deep and Transformer-based trackers, while earlier works are included when necessary for historical context. We adopted explicit inclusion/exclusion rules. Included papers are those that (i) propose or substantially improve a visual single-object tracker applicable to UAV scenarios, or (ii) provide UAV-relevant evaluation, analysis, or deployment discussion, especially with experiments on UAV benchmarks or aerial videos. Excluded papers are those that focus purely on aerial detection without tracking, multi-object tracking without a single-object tracking setting, non-visual tracking (e.g., radar-only) without visual target tracking components, or generic tracking papers without

UAV-related evaluation or discussion. To reduce omission bias, we further performed backward/forward snowballing from highly cited UAV tracking papers and recent surveys, and we cross-checked representative methods in each category to ensure balanced coverage across correlation-filter, Siamese, and Transformer-based trackers. Finally, each included work was mapped to the five UAV challenges in our taxonomy, and key information (core idea, computational characteristics, and reported evaluation settings) was extracted to support the quantitative and qualitative comparisons presented in later sections.

2.1 Major Categories and Development of Existing Object Tracking Methods

The development of visual object tracking methods can be broadly divided into three stages: traditional methods, deep learning-based ones, and emerging ones. Research efforts at different stages exhibit significant differences in model assumptions, feature representation capabilities, and adaptability to complex scenarios, leading to varying degrees of suitability for UAV target tracking tasks.

2.1.1 Traditional Methods

Early studies on object tracking mainly relied on classical state estimation and probabilistic modeling theories, such as Kalman filtering [16], the MeanShift algorithm [17], as well as correlation filter-based methods represented by MOSSE [37], KCF [18], and SRDCF [38]. These methods typically depend on hand-crafted features and linear update strategies, resulting in low computational complexity and strong real-time performance. Consequently, they were widely adopted in early real-time tracking systems.

In UAV scenarios, traditional methods can partially satisfy the real-time requirements under limited onboard computational resources, especially in scenes with simple backgrounds and smooth target motion. However, due to the prevalence of small target sizes, drastic scale variations, rapid viewpoint changes, and severe background interference in aerial videos, traditional methods often struggle to maintain stable performance in complex environments, and are prone to model drift and tracking failure. Moreover, correlation filter-based methods usually rely on local search regions and linear assumptions, which limit their ability to handle long-term occlusions and fast maneuvering targets [8].

2.1.2 Deep Learning-Based Methods

With the development of deep learning, object tracking methods based on Siamese networks have significantly enhanced feature representation capabilities and environmental adaptability. By learning the similarity mapping between the template and the search region, Siamese networks achieve favorable real-time performance while ensuring high accuracy, gradually becoming the mainstream solution for UAV object tracking. In UAV applications, Siamese networks and their improved variants can more effectively cope with complex background interference and scale variations, achieving significantly superior performance compared with traditional methods on multiple aerial benchmark datasets. In addition, discriminative deep tracking methods, such as ATOM [39], DiMP and its improved variants [24,40], further enhance tracking accuracy and robustness in complex scenarios by introducing online optimization and target discrimination modeling mechanisms. In this paper, the term optimization refers to standard learning or inference procedures used in tracking, such as online model adaptation and deployment-oriented compression, rather than meta-heuristic search methods.

Despite their clear performance advantages, deep learning-based methods usually involve large model sizes and high computational costs, which pose considerable challenges for deployment on onboard UAV platforms. How to reduce model complexity while maintaining tracking accuracy remains a critical issue for deep tracking methods in UAV scenarios.

2.1.3 Emerging Methods

In recent years, Transformer architectures have been increasingly introduced into visual tracking and have shown strong potential in UAV target tracking, mainly due to their ability to model global context and long-range dependencies. By using self-attention to capture global interactions between the target and background, Transformer-based trackers can improve discriminability under UAV difficulties such as small targets, cluttered scenes, and rapid motion. Representative UAV-oriented routes go beyond listing methods and can be grouped by their technical focus. HiFT enhances target representation through hierarchical feature interaction across multiple scales, which is beneficial for low-resolution targets and scale changes in aerial views [25]. TCTrack emphasizes temporal-context modeling to improve robustness under fast motion and viewpoint changes, and it reports real-time feasibility on UAV-related platforms [41]. These studies reflect a transformer-enhanced design route that strengthens feature interaction and temporal consistency for UAV tracking.

More recent works further push efficiency and robustness for practical deployment. Aba-ViTrack reduces inference cost by adaptively suppressing background tokens while maintaining competitive accuracy, which directly addresses the compute-robustness trade-off on UAV platforms [42]. AVTrack introduces adaptive computation and view-invariant representation learning to better cope with UAV viewpoint variation and camera motion [43]. ORTrack targets frequent occlusion in aerial scenes by learning occlusion-robust representations and additionally provides a compact distilled variant for improved efficiency [32]. Meanwhile, lightweight designs remain essential for onboard real-time tracking. LiteTrack pursues efficiency through structural pruning and asynchronous feature processing, demonstrating that careful lightweight design can deliver practical speed on edge devices [44]. Although Transformer-based methods continue to advance, the overall model complexity and inference cost are still major barriers, and efficient onboard deployment under compound UAV challenges remains an open and important direction.

To intuitively illustrate the evolution of different tracking paradigms, Fig. 3 presents a chronological timeline of representative object tracking algorithms from 2010 to 2025. The timeline highlights a significant paradigm shift from early Correlation Filter methods, typified by MOSSE and KCF, to the Deep Siamese Network framework established by SiamFC and SiamRPN. In the post-2021 era, Transformer-based architectures such as HiFT, TCTrack, and ORTrack have emerged as a dominant trend due to their superior global modeling capabilities. Notably, driven by the computational constraints of UAV platforms, lightweight algorithms marked in green, exemplified by LightTrack and LightFC, have garnered increasing attention, reflecting the research trajectory towards balancing tracking precision with onboard computational efficiency.

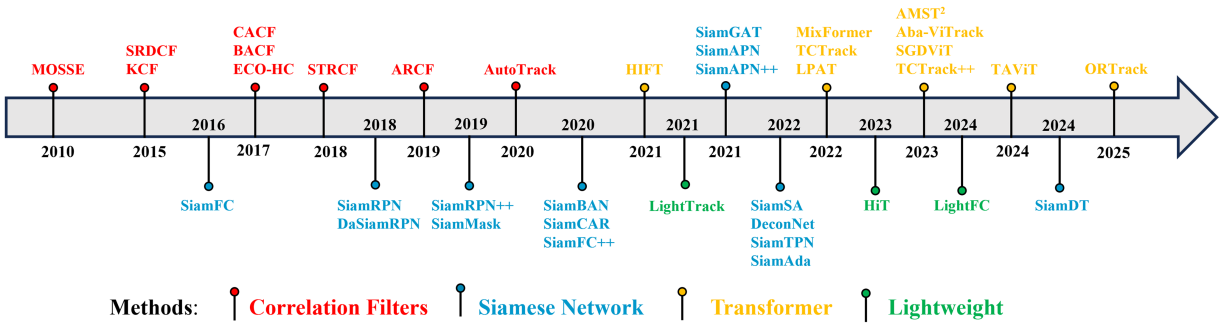

Figure 3: Development timeline of object tracking methods.

2.2 Research Gaps in Existing Survey Works

Survey studies on visual object tracking have achieved certain progress. However, notable limitations remain, particularly in the domain of UAV target tracking. Early survey works mainly focused on generic visual tracking tasks and provided systematic reviews of mainstream frameworks and evaluation benchmarks, such as discriminative correlation filters and Siamese networks [45–47]. Nevertheless, these surveys paid insufficient attention to challenges commonly encountered in UAV aerial scenarios, including small targets, rapid viewpoint changes, drastic scale variations, and onboard computational and real-time constraints, with related discussions often confined to benchmark evaluations or isolated phenomena [10].

In recent years, several studies have begun to specifically address UAV target tracking through survey analyses [6–9,48,49]. For instance, some works systematically reviewed the development of Siamese network-based tracking methods for UAV applications and provided relatively comprehensive analyses [6]. Other studies focused on correlation filter-based tracking algorithms in UAV scenarios and presented experimental evaluations and comparisons [8]. In addition, some surveys summarized research progress in UAV target detection and tracking from a deep learning perspective [7,9], offered classifications and systematic reviews for more generalized UAV moving target tracking tasks [48], or discussed issues related to real-time applications from the perspectives of remote sensing or UAV-based detection [49]. These works provide valuable references for understanding the development of UAV target tracking.

Although several of these works also discuss UAV difficulties, their organizing axes are not the same. In particular, [6,8] mainly organize the literature around a single method category, where the discussion is conducted within correlation-filter trackers or within Siamese trackers. In these studies, challenge attributes are commonly used as evaluation labels or as auxiliary analysis dimensions, rather than as the primary taxonomy that unifies multiple method categories. Surveys such as [7,9] provide broader coverage of UAV tracking, but their organizations are typically driven by system pipelines, task scope, or perception modules. This is effective for presenting an overall picture, yet it does not explicitly construct a unified mapping from UAV visual difficulty patterns to tracker design mechanisms across different method categories. In addition, [48] categorizes UAV moving target tracking from a requirement and task perspective, and [49] emphasizes real-time aerial perception from remote sensing and deployment viewpoints. Both are valuable for system understanding, but they are not designed to organize single-object UAV tracking methods under a unified set of UAV-specific visual challenges.

However, existing surveys still suffer from several limitations, which can be summarized as follows:

(1) **Limited coverage of methods:** Most surveys concentrate on a single category of methods (e.g., Siamese networks or correlation filters), lacking systematic comparisons and unified analyses across different technical paradigms;
(2) **Insufficiently systematic summarization of UAV-specific challenges:** Issues such as night-time low-light conditions, complex background interference, and multi-modal sensor fusion are often discussed in a fragmented manner, and a unified analytical framework has yet to be established;
(3) **Limited attention to emerging directions:** With the rapid development of Transformers, multi-modal tracking, and lightweight deployment techniques, systematic summaries of these emerging directions remain insufficient in existing survey works.

To further illustrate the differences in coverage among existing surveys, Table 1 presents a comparative analysis of representative survey papers from multiple dimensions, including application scenarios, research targets, and task types. Compared with existing works, our survey uses five core UAV challenges as the primary organizing axis and treats correlation-filter-based, Siamese-based, and Transformer-based trackers as complementary views within each challenge. This design directly supports cross-category comparison

under the same difficulty patterns and enables cross-challenge applicability analysis by examining whether a technical mechanism can generalize across multiple UAV challenges and what robustness-efficiency trade-offs it implies for practical onboard deployment. Therefore, the novelty of our classification does not lie in simply adding another list of categories, but in providing a unified challenge-to-mechanism analytical lens. Under this lens, we organize and interpret major tracker categories, including correlation filters, Siamese networks, and Transformer-based trackers, under the same set of five UAV-specific challenges. This improves coverage completeness and connects quantitative comparisons and failure cases more directly to deployment constraints and robustness–efficiency trade-offs. To the best of our knowledge, existing UAV-oriented surveys have not explicitly established a unified framework that simultaneously organizes and compares major tracker categories under the same five challenges, and our challenge-driven taxonomy is proposed to address this gap.

Table 1: Comparison of existing surveys (✓: Explicitly discussed, ◐: Partially discussed, ✗: Not involved).

Survey	UAV-Specific	Challenge Analysis	CF-Based	Siamese-Based	Transformer-Based
[6]	✓	✓	✗	✓	◐
[7]	✓	✓	◐	✓	◐
[8]	✓	✓	✓	◐	✗
[9]	✓	✓	✓	✓	✗
[10]	✓	◐	◐	✓	✗
[45]	✗	◐	✓	✓	✗
[46]	✗	◐	◐	✓	◐
[47]	✗	◐	✗	◐	✓
[48]	✓	✓	✓	✓	✗
[49]	✓	✓	✗	◐	✗
This article	✓	✓	✓	✓	✓

3 UAV Target Tracking Methods

Prior surveys can be organized from different viewpoints. Some categorize methods by model families, such as correlation-filter trackers, Siamese trackers, and Transformer trackers. Others organize by pipeline components, such as feature representation, similarity matching, model update, and re-detection. There are also task-oriented views, such as short-term vs. long-term tracking, or single-modal vs multi-modal tracking. These perspectives are helpful for describing the evolution of algorithms, but they do not always explain why a tracker succeeds or fails under UAV deployment constraints. In UAV scenarios, platform motion and aerial imaging repeatedly produce a set of dominant difficulty patterns. They include small targets with strong scale changes, rapid motion with motion blur, cluttered backgrounds with frequent occlusions, night-time or low-light conditions, and strict real-time and efficiency constraints on onboard hardware. These difficulties directly shape failure modes and practical usability in the field. In addition, UAV benchmarks are often analyzed using challenge attributes such as low resolution, occlusion, illumination variation, and camera motion, which indicates that performance is closely tied to scenario difficulties rather than method names alone. Therefore, we adopt a challenge-driven taxonomy as the primary organizing principle and use model-family categories as a complementary view inside each challenge. This design links technical mechanisms to UAV-specific difficulties, supports unified interpretation of quantitative results and failure cases, and highlights deployment trade-offs between robustness and efficiency.

The design of UAV target tracking methods is highly dependent on the practical challenges encountered in real-world applications. Unlike generic visual object tracking, UAV platforms exhibit significant differences in imaging viewpoints, target scale distributions, motion patterns, and onboard computational resources, which leads to notable variations in the applicability and effectiveness of existing tracking methods. To facilitate a more systematic understanding of prior studies, this section analyzes UAV target tracking methods from the perspective of the core challenges inherent to UAV scenarios.

Specifically, this study analyzes UAV target tracking methods by focusing on five representative challenges commonly encountered in UAV scenarios, including small targets with drastic scale variations, rapid platform motion and motion blur, complex backgrounds with frequent occlusions, nighttime and low-light conditions, as well as lightweight design and real-time requirements under constrained onboard computational resources.

Based on these dimensions, Table 2 provides a systematic summary of representative correlation filter (CF), Siamese network (SN), and Transformer-based tracking methods in recent years, comparing their research emphasis and capability coverage across different challenge aspects. It can be observed that traditional correlation filter-based methods generally exhibit high computational efficiency and favorable real-time performance, but their ability to handle complex backgrounds, severe occlusions, and nighttime or low-light conditions remains limited. In contrast, Siamese network-based approaches leverage deep feature representations and end-to-end similarity learning mechanisms, achieving notable performance improvements in scenarios involving small targets, scale variations, and background interference, and have gradually become the mainstream technical paradigm in UAV target tracking. More recently, Transformer-based trackers employ global modeling and cross-frame feature interactions to demonstrate stronger representation capability in challenging scenarios such as complex backgrounds, occlusions, and rapid motion; however, their computational cost and deployment efficiency still require further optimization.

Table 2: Comparison of UAV Target Tracking Methods (✓: Explicitly discussed, ◐: Partially discussed, ✗: Not involved). Metrics: (1) Small Targets & Scale Variation, (2) Platform Motion & Motion Blur, (3) Complex Background & Occlusion, (4) Nighttime & Low-light, (5) Lightweight & Real-time.

Method	Year	Category	(1)	(2)	(3)	(4)	(5)
KCF [18]	2015	CF	◐	◐	✗	✗	✓
SRDCF [38]	2015	CF	◐	◐	◐	✗	✗
BACF [50]	2017	CF	◐	◐	✓	✗	◐
STRCF [51]	2018	CF	◐	✓	◐	✗	◐
ARCF [52]	2019	CF	✓	◐	✓	✗	✓
AutoTrack [53]	2020	CF	✓	✓	✓	✗	✓
BiCF [54]	2020	CF	◐	◐	✓	✗	✓
IBRI [55]	2021	CF	◐	◐	✓	✗	✓
SiamFC [19]	2016	SN	◐	◐	◐	✗	✓
DaSiamRPN [56]	2018	SN	◐	✓	✓	✗	✓
SiamRPN++ [57]	2019	SN	✓	✓	◐	✗	◐
SiamAPN [22]	2021	SN	✓	✓	◐	✗	✓
SiamAPN++ [23]	2021	SN	✓	✓	◐	✗	✓
LightTrack [58]	2021	SN	◐	◐	✗	✗	✓
SiamSA [12]	2022	SN	✓	◐	◐	✗	◐
UDAT [59]	2022	SN (Night)	◐	◐	◐	✓	◐
CDT [60]	2023	SN (Night)	◐	◐	◐	✓	◐

(Continued)

Table 2 (continued)

Method	Year	Category	(1)	(2)	(3)	(4)	(5)
HiFT [25]	2021	Transformer	✓	✓	✓	✗	◐
TCTrack [41]	2022	Transformer	◐	✓	✓	✗	◐
HiT [61]	2023	Transformer	◐	◐	◐	✗	✓
SGDViT [26]	2023	Transformer	✓	✓	◐	✗	◐
AVTrack [43]	2025	Transformer	✓	✓	✓	✗	◐
ORTrack [32]	2025	Transformer	◐	◐	✓	✗	◐

It should be noted that this comparison does not constitute a quantitative performance evaluation of tracking algorithms, but rather summarizes the design focus and application scope of different technical paradigms from a methodological perspective. Through this challenge-oriented horizontal analysis, the strengths and limitations of different tracking approaches in UAV scenarios can be more intuitively understood, thereby providing an overall reference framework for the subsequent challenge-specific discussions.

3.1 Tracking Methods for Small Targets and Drastic Scale Variations

UAVs typically operate at high altitudes, where targets appear small in size, low in resolution, and with limited texture information in captured images. Meanwhile, target scales frequently change due to variations in UAV altitude or target motion. These factors significantly degrade the effectiveness of traditional local features and fixed-scale search strategies, making small targets and drastic scale variations one of the most prominent challenges in UAV target tracking. Small targets usually exhibit sparse features and are easily overwhelmed by background clutter, while rapid scale changes hinder the adaptability of model templates, thereby increasing the risk of tracking drift and target loss.

To address issues related to small targets and scale variations, researchers have mainly focused on scale-adaptive modeling [62], multi-scale feature fusion [63], small-target feature enhancement [22], context modeling [41], and attention mechanisms [12]. Early methods primarily introduced scale estimation mechanisms within correlation filter frameworks. DSST [62] explicitly builds a scale space by sampling multiple scaled candidates around the predicted target location and learning a dedicated scale filter. During inference, it evaluates responses across the scale pool and selects the scale with the strongest response, which provides a direct mechanism for handling continuous scale changes. fDSST [64] follows the same scale-space principle but improves efficiency so that scale estimation can be executed more frequently, reducing lag when the target size changes over time. However, in UAV aerial tracking, scale changes are often abrupt and large. When the target becomes extremely small, the response surface becomes noisy and multi-peaked, so discrete scale pools can be misled by background clutter, and scale estimation errors can quickly propagate into template updates.

Subsequently, methods such as ARCF [52], AutoTrack [53] were proposed to better accommodate UAV-specific characteristics by strengthening robustness of representation and reducing drift under scale variations. ARCF [52] improves scale robustness by emphasizing reliability in the correlation learning process, which down-weights unreliable spatial regions and unstable channels. This is important for small targets because a small bounding box is easily contaminated by surrounding background pixels. AutoTrack [53] further improves stability by refining the learning and update behavior of the appearance model. Its core idea is to reduce drift accumulation under abrupt scale and appearance changes by making the tracker less sensitive to short-term noise and less aggressive in updating uncertain templates. These approaches typically

rely on scale pyramids on feature maps and incorporate additional constraints during optimization, which makes the estimated scale less sensitive to local noise and improves stability in cluttered backgrounds.

With the advancement of deep learning frameworks, Siamese networks have become the mainstream solution for small-target tracking. The SiamAPN series [22,23] introduces anchor proposal mechanisms to significantly improve small-target localization accuracy while maintaining high computational efficiency. In SiamAPN [22], multi-scale anchors generate candidate regions of different sizes, and the tracker performs classification and regression on these candidates. This provides an explicit measure for scale variation, because the predicted box is selected from multiple size hypotheses rather than being forced into a fixed-scale matching peak. SiamAPN++ [23] strengthens the aggregation and interaction of features used for scoring and regression, which improves localization stability when the target scale changes rapidly and the visual evidence is weak. SiamSA and its extended variants [12] provide another concrete measure by introducing scale-channel attention. They dynamically reweight features from different scales and channels according to the current observation, making the tracker focus on more informative scale cues and suppress noisy responses when the target is tiny or undergoing drastic scale changes.

To address insufficient feature representation for extremely small targets, SmallTrack [65] enhances target discriminability through wavelet pooling, graph-structured modeling, and contextual information fusion. Wavelet pooling [66] strengthens edge and structure cues that are relatively stable under low resolution, which directly mitigates the loss of high-frequency texture in tiny targets. Graph-based modeling captures relationships among local target cues, so the decision is supported by structured feature interactions rather than a single fragile point. Context fusion further complements sparse target appearance by incorporating surrounding information that correlates with target identity and motion, improving robustness when the target occupies only a few pixels.

In recent years, Transformer architectures have also been introduced into small-target tracking research. HiFT [25] adopts hierarchical feature interaction to perform multi-scale contextual modeling, which provides a concrete mechanism to connect fine details and semantic context for scale robustness. AbaViTrack [42] enhances small-target tracking by suppressing background interference and strengthening target-focused representation, which is especially helpful when the target signal is weak and easily confused with clutter. AVTrack [43] further improves stability under viewpoint variation that often co-occurs with scale changes, by combining view-invariant modeling with cross-frame attention mechanisms. Furthermore, some studies attempt to fuse multi-modal information, such as infrared or thermal imaging [14,67], with visual features. This provides complementary cues when visible-light texture is insufficient, helping alleviate information loss for small targets under low contrast or degraded imaging conditions.

Overall, methods for small targets and drastic scale variations evolve by making scale estimation more explicit, strengthening multi-scale representation, enhancing weak target cues, and controlling updates under uncertain observations. These concrete measures collectively reduce drift and target loss in UAV aerial tracking, while maintaining feasibility under onboard constraints.

3.2 Tracking Methods for Rapid Platform Motion and Motion Blur

The high-speed flight, attitude changes, and external airflow disturbances of UAVs often lead to significant viewpoint variations and motion blur, making it difficult to stably model target appearance features and motion patterns. In high-altitude or high-speed aerial imaging, targets may exhibit large inter-frame displacements, causing traditional trackers based on local template matching to easily lose the target. Meanwhile, motion blur results in the degradation of target texture information, which further complicates feature extraction.

To address rapid motion, researchers have mainly explored directions such as temporal modeling, motion compensation, multi-frame information fusion, and blur-robust feature learning. Traditional correlation filter-based methods, such as STRCF [51] and CPCF [68], enhance adaptability to fast-moving targets by introducing temporal consistency constraints and response stability modeling. STRCF [51] explicitly constrains the filter update to be temporally smooth, so that the learned model does not change abruptly between adjacent frames. This is particularly important under fast motion, because rapid viewpoint shifts and partial blur can easily generate unstable responses that would otherwise corrupt online updates. CPCF [68] further emphasizes response stability by suppressing unreliable response peaks and encouraging more consistent correlation responses over time, which reduces the risk of drifting to distractors when the target displacement becomes large. By combining historical responses with current features for dynamic template updating, these methods reduce sensitivity to transient motion anomalies or frame skipping. However, when fast motion persists for a long period, small update errors can accumulate, and purely local correlation responses may still become unreliable, leading to gradual performance degradation.

Within deep learning frameworks, discriminative tracking methods such as ATOM [39], DiMP and its improved variants [24,40], enhance robustness to rapid motion through online optimization and target-background discrimination modeling. ATOM [39] separates the tracking process into a classification component and a localization component, and it learns a discriminative classifier online via optimization. This design helps under rapid motion because the classifier can adapt to fast appearance changes and background shifts, while the localization component provides a more stable bounding-box estimate when the target moves quickly. DiMP [40] further improves this idea by learning a model predictor that produces a more effective classifier during online optimization, which strengthens target-background separation when blur weakens low-level textures. The probabilistic variant [24] enhances robustness by modeling uncertainty in the learning and update process, which makes the tracker less likely to overfit to noisy observations caused by fast motion and blur. In addition, some approaches integrate optical flow or motion estimation modules to predict target positions, enabling trackers to compensate for displacement in advance and improve stability across consecutive frames.

To specifically address motion blur in UAV scenarios, several methods explicitly incorporate cross-frame temporal context modeling. TCTrack [41] strengthens robustness to rapid viewpoint changes and blurred scenes by modeling temporal context information across frames, so that tracking decisions rely on accumulated cues rather than a single blurred frame. Building upon this framework. BDTrack [69] introduces a dynamic early-exit mechanism and blur-aware design. It adjusts computation according to tracking difficulty, and activates stronger processing when blur is severe, while exiting earlier when confidence is sufficient, thereby reducing computational cost without sacrificing accuracy under challenging conditions. In addition, some studies attempt to combine motion compensation with image deblurring techniques to extract clearer features for tracking, which mitigates feature degradation caused by motion blur.

In practical applications, rapid platform motion and motion blur often co-occur with small targets and complex backgrounds. Consequently, multi-modal feature fusion [70], multi-scale temporal modeling [41], and dynamic template updating strategies [71] have become key techniques for addressing high-speed motion and motion blur. Multi-modal fusion [70] provides complementary cues when visible textures are degraded by blur, multi-scale temporal modeling [41] strengthens cross-frame consistency under large displacement, and dynamic template updating [71] helps prevent corrupted updates when observations are unreliable. Future research directions include incorporating more refined motion prediction modules, leveraging Transformers for global temporal modeling [72], and integrating augmented reality or multi-sensor information to improve tracking accuracy and real-time performance on high-speed UAV platforms.

3.3 Tracking Methods for Complex Backgrounds and Frequent Occlusions

Aerial scenes typically contain highly complex background structures and numerous visually similar distractors, such as buildings, vegetation, water surfaces, and roads. These background elements may share similar color, texture, or shape characteristics with the target, making trackers prone to mismatching and model contamination. Moreover, scene components such as buildings and trees often cause partial or long-term occlusions, leading to the temporary disappearance of target features in the image and significantly increasing the difficulty of maintaining tracking continuity. Under such challenging conditions, traditional template-based or single-frame feature-based tracking methods are highly susceptible to background interference, resulting in degraded tracking accuracy or complete target loss.

To mitigate complex background interference, many approaches introduce saliency modeling, background suppression, and target consistency constraints. For instance, BiCF [54] effectively reduces the influence of background distractors by modeling response inconsistency and incorporating interference suppression mechanisms. IBRI [55] combines multi-scale feature representation with interference suppression strategies to enhance target-background discrimination in cluttered environments. These methods dynamically reweight target and background responses, enabling the tracker to maintain accurate localization of the true target even under severe background interference.

In deep Siamese networks, DaSiamRPN [56] demonstrates higher robustness in complex scenarios by explicitly modeling distractor objects and local background information. Such methods typically employ region proposal networks to generate multiple candidate regions and enhance target representations through feature fusion and contextual modeling, thereby reducing mismatches caused by background interference. Compared with traditional approaches, deep networks can automatically extract discriminative target features via convolutional representations, improving the recognition of small targets in cluttered aerial backgrounds.

To address occlusion challenges, several studies incorporate re-detection and candidate association mechanisms. KeepTrack [73] maintains strong tracking continuity in long-term occlusion scenarios by leveraging multi-candidate association and segmentation-assisted strategies. RTS [74] integrates segmentation cues with multi-candidate matching to rapidly recover target positions after occlusion events. In UAV scenarios, ORTrack [32] and Query-Guided Redetection [75] methods further exploit the global modeling capability of Transformers, enabling feature reasoning through global interactions between the target and candidate regions, which significantly enhances recovery performance under severe occlusion conditions.

3.4 Tracking Methods for Nighttime and Low-Light Conditions

Nighttime and low-light conditions represent typical extreme scenarios in UAV target tracking. The primary challenges include severe imaging noise, low contrast, loss of color information, and unclear target textures. Under low-light conditions, the visual distinction between targets and backgrounds is significantly reduced, making it difficult for conventional RGB-based trackers to extract discriminative features and leading to substantial performance degradation. Furthermore, nighttime missions are often accompanied by motion blur and low-resolution imagery, which further complicate accurate target localization.

To address nighttime tracking challenges, a variety of specialized methods have been proposed. ADTrack [15] adopts an all-time modeling strategy to achieve a unified tracking framework for both daytime and nighttime scenarios, maintaining stable performance across varying illumination conditions. HighlightNet [76] enhance target visibility and feature discriminability in low-light images through illumination enhancement and feature compensation mechanisms. These methods typically integrate enhancement

modules during the feature extraction stage, performing brightness amplification and noise suppression to enable the tracker to capture salient target features under poor lighting conditions.

UDAT [59] employs an unsupervised domain adaptation strategy to compensate for the domain shift between daytime and nighttime scenes, alleviating feature distribution discrepancies caused by illumination variations. Within Transformer-based frameworks, CDT [60] introduce nighttime-aware modules and cascaded denoising mechanisms to improve robustness under extreme illumination conditions. By leveraging global feature interactions and temporal modeling, these approaches enhance the capture of weak target features, enabling reliable tracking performance even in nighttime environments.

In addition, methods such as DCPT [77] and MambaNUT [78] further combine prompt learning, diffusion modeling, and sequence modeling to perform multi-level enhancement of nighttime target features. In practical applications, these approaches effectively improve target detection and localization accuracy in low-contrast and high-noise environments while maintaining computational efficiency. Nighttime tracking methods commonly integrate multi-frame information and feature compensation strategies, leveraging historical frame information to reinforce current-frame representations and ensure sustained target trackability under low-light conditions.

3.5 Lightweight and Real-Time Tracking Methods for Onboard Resource-Constrained Platforms

Due to payload and battery limitations, UAVs are typically equipped with computational platforms of limited processing capability, which directly restricts the deployment of high-complexity models during real-world flight. Consequently, target tracking algorithms for UAV applications must strike a careful balance among tracking accuracy, model complexity, and real-time performance. Lightweight model design and efficient inference have therefore become critical research directions in UAV target tracking. These approaches aim to maintain robustness against small targets, complex backgrounds, and motion blur while satisfying the real-time computational constraints of embedded platforms.

Within Siamese network frameworks, LightTrack [58] achieve efficient deployment on embedded platforms through neural architecture search and lightweight network design strategies. By reducing convolutional parameters and adopting techniques such as depthwise separable convolutions and channel pruning, these methods compress originally complex networks into lightweight models suitable for mobile and embedded devices, while preserving sensitivity to small and multi-scale targets.

Several studies optimize existing tracking models through pruning, quantization, low-bit feature representation, and feature compression strategies. For example, improved versions of SiamCAR [79] employ channel pruning and weight-sharing mechanisms to reduce redundant computation, while FERMT [80] significantly lowers computational cost on embedded platforms via quantization and sparsification techniques. These optimization methods not only reduce model storage requirements but also decrease inference latency, allowing UAVs to sustain real-time tracking performance during long-duration flight missions.

In the domain of Transformer-based tracking, HiT [61] and HCAT [81] enhance real-time capability by introducing hierarchical attention and efficient cross-attention mechanisms. These approaches decompose global feature interactions into hierarchical stages, transforming high-dimensional attention computations into lower-dimensional subspace operations. Combined with local window attention to reduce redundant computation, they achieve a favorable balance between tracking accuracy and efficiency. In addition, works such as E.T.Track [82] and LightFC [83] further improve computational efficiency from the perspectives of training strategies and feature correlation modeling. For example, E.T.Track optimizes the inference process through feature selection and lightweight attention modules, while LightFC leverages sparse convolution and lightweight feature fusion strategies to achieve high frame rates on UAV embedded platforms.

Beyond network architecture optimization, some studies integrate hardware acceleration and inference optimization strategies to meet real-time UAV tracking requirements [84]. Models can be quantized and accelerated on NPU, GPU, or FPGA to reduce memory access overhead and computational latency, while dynamic resolution adjustment and multi-frame aggregation strategies further decrease computational burden. These approaches demonstrate favorable real-time performance and stability in practical UAV missions, maintaining high tracking accuracy even in scenarios involving complex backgrounds and small targets.

Through the adoption of lightweight modeling and efficient inference strategies, UAV tracking models are able to achieve real-time, high-precision target tracking under onboard computational constraints, while maintaining robust performance in challenging conditions such as small targets, rapid motion, complex backgrounds, and low-light environments. This line of research not only facilitates the practical deployment of tracking algorithms in real UAV systems but also provides a foundation for future studies on optimizing high-performance tracking models in embedded environments.

4 Comparison and Evaluation

This section presents a comparative evaluation of the aforementioned tracking methods across multiple UAV target tracking benchmarks. Considering the significant differences among UAV scenarios in terms of imaging altitude, viewpoint variation, target scale, and motion patterns, five widely used benchmarks—UAV123 [85], UAV123@10fps [85], UAVDT [13], UAV20L [85], and DTB70 [11]—are selected for experimental validation. Quantitative comparisons are conducted from three perspectives: tracking precision, success rate, and real-time performance. It is worth noting that, when summarizing and comparing performance across trackers, we primarily adopt the results reported in the original papers so that the evaluation protocol and dataset usage remain consistent with the authors' settings. Unless otherwise stated, these numbers are not reproduced by us. When such information is available, we also provide the testing setup and environment described in the original works, so that readers can interpret the reported results with awareness of potential differences caused by implementation details and hardware conditions.

4.1 Benchmark Datasets

We evaluate trackers on five UAV benchmarks: UAV123, UAV123@10fps, UAV20L, DTB70, and UAVDT. These datasets cover small targets and scale variation, fast motion with motion blur, long-term tracking, significant camera motion, and complex real-world conditions with occlusion.

UAV123 is a representative aerial single-object tracking benchmark that contains 123 fully annotated video sequences with 112,578 frames, covering diverse scenes and objects captured from low-altitude UAV perspectives. It provides a broad evaluation basis for typical UAV challenges such as scale variation, viewpoint changes, and background clutter.

UAV123@10fps is constructed by downsampling UAV123 from 30 frames per second to 10 frames per second, which imitates discontinuous aerial video sampling and increases the difficulty caused by larger inter-frame motion. This benchmark is particularly useful for evaluating robustness under fast motion and motion blur.

UAVDT is a UAV benchmark focusing on complicated real-world conditions such as weather, altitude, camera view, vehicle type, and occlusion. Its single-object tracking subset includes 50 sequences with 37,084 frames. It provides complementary evaluation for cluttered scenes and frequent occlusions that commonly occur in UAV applications.

UAV20L is designed for long-term aerial tracking and consists of the 20 longest sequences formed by merging short subsequences, which makes target location changes between frames larger and more irregular. It is therefore used to assess tracking stability and recoverability over long durations.

DTB70 consists of 70 sequences totaling 15,777 frames, and it aims to address the difficult issue of significant camera motion under various extreme conditions. This benchmark is suitable for analyzing performance under severe platform motion and background changes.

4.2 Evaluation Metrics

Three commonly adopted evaluation metrics in the UAV tracking community are employed in this study:

(1) **Precision:** which measures the center location error (CLE) between the predicted target position and the ground truth, with a threshold of 20 pixels;

(2) **Success rate:** which is based on the intersection-over-union (IoU) between the predicted bounding box and the ground truth. The success plot is obtained by computing success rates under varying IoU thresholds, and the area under the curve (AUC) is used as the final metric;

(3) **Real-time performance:** which is evaluated by the tracking speed in frames per second (FPS) to assess the feasibility of deploying the algorithm under onboard computational constraints. Runtime speed is highly dependent on the testing platform, so FPS values reported in different papers are not directly comparable like accuracy metrics. In this work, the FPS of CF-based trackers mainly comes from evaluations on desktop GPU platforms, while the FPS of Siamese-based trackers is mainly summarized from onboard Jetson platforms, and is therefore used only to indicate approximate runtime scale rather than for strict cross-platform ranking.

4.3 Experimental Results and Comparative Analysis

To better support UAV deployment and method selection, we further analyze the typical failure triggers and limitations of different tracker categories from a mechanism-oriented perspective. For correlation filter trackers, failures are frequently triggered by rapid scale and appearance changes, because the filter cannot adapt in time and may learn incorrect target information, which eventually leads to drift. Occlusion is another common cause, since partial occlusion corrupts the learned template and full occlusion makes it difficult to re-acquire the target when it re-enters the view. Low-resolution targets further aggravate this issue by providing insufficient discriminative cues, and poor lighting often prevents the filter from learning representative features, especially when combined with scale variation or similar objects. For Siamese trackers, the main failures are closely related to limited effective resolution, occlusion, and viewpoint induced deformation. When the target and background are both low-resolution, feature extraction becomes unreliable and the tracker struggles to distinguish foreground from clutter. Occlusion leads to incomplete target features or temporary disappearance, and the viewpoint and relative motion between camera and target cause deformation that accumulates across frames, which can produce erroneous matching and loss of the target. Transformer based trackers can better exploit global context and long-range interactions, but they still face practical failure risks from high computational cost and deployment constraints on edge platforms. This makes it difficult to maintain real-time operation while preserving robustness under small targets, clutter, and long occlusion, and it also explains why a fully stable all-weather onboard deployable solution has not yet been achieved in UAV tracking. These failure patterns provide a more balanced understanding of current methods and motivate future designs that jointly address robustness and efficiency under the five UAV challenges.

To provide an intuitive analysis of the performance differences among various tracking methods in complex UAV scenarios, representative failure cases are visualized in Fig. 4. Six representative sequences are selected from the DTB70 dataset, namely Animal1, Car2, Horse1, Soccer1, Surfing03, and MountainBike1. These sequences are not randomly chosen, but correspond to different UAV tracking challenge attributes. Specifically, Animal1 mainly involves scale variation, aspect ratio variation, and surrounding similar objects; Car2 is characterized by fast camera motion, motion blur, and similar-object interference; Horse1 mainly reflects occlusion, background clutter, and surrounding similar objects; Soccer1 contains fast camera motion, in-plane rotation, motion blur, and surrounding similar objects; Surfing03 emphasizes aspect ratio variation, fast camera motion, in-plane rotation, and motion blur; and MountainBike1 mainly reflects occlusion and surrounding similar objects. Therefore, these visualized cases collectively cover several core UAV challenges, including scale variation, platform motion and motion blur, and complex backgrounds with occlusion or distractors. Bounding boxes in different colors represent the tracking results of **SiamAPN**, **HiFT**, **TCTrack**, **AutoTrack**, **LightTrack**, and the green boxes denote the ground-truth target locations.

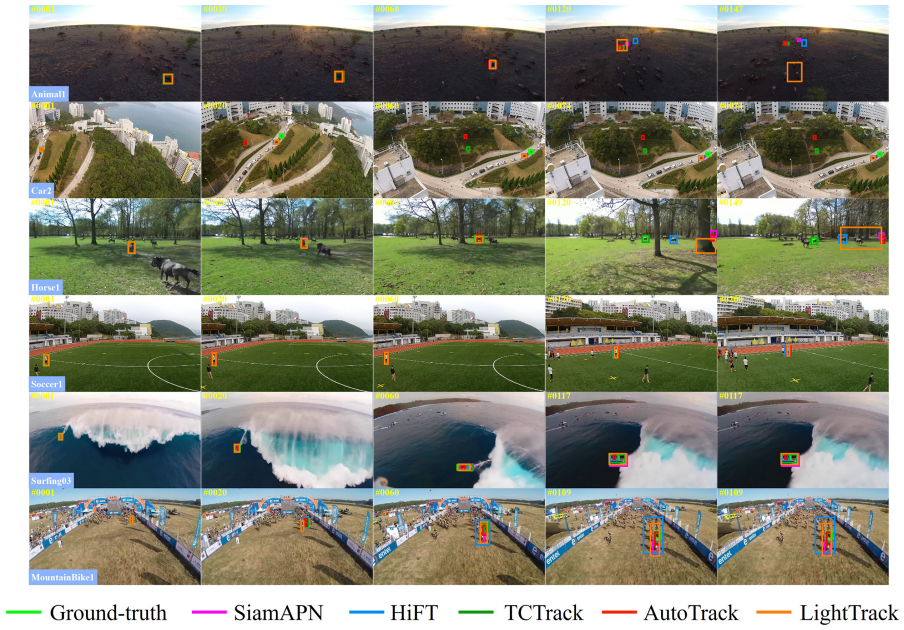

Figure 4: Comparison of typical failure cases of different trackers. The main challenges illustrated in these figures are scale variation, aspect ratio variation, occlusion, fast camera motion, in-plane rotation, out-of-plane rotation, out of view, background clutter, similar objects around, motion blur.

In addition, Fig. 5 shows tracker failure cases in nighttime low-light scenes. The selected sequences mainly feature weak target texture, uneven lighting distribution, noticeable glare from car headlights or streetlights, and a low signal-to-noise ratio. In this figure, the green boxes denote the ground-truth target locations, while the orange, magenta, red, and blue boxes correspond to **LightTrack**, **SiamFC++**, **AutoTrack**, and **SiamAPN**, respectively. As can be seen, LightTrack shows the most pronounced failure in S0302 and S0308, where its predicted boxes expand rapidly and drift toward bright background regions or salient illuminated structures, indicating that its lightweight representation is more vulnerable to severe illumination interference and distractor attraction in low-light scenes. SiamFC++ generally preserves the target more stably than LightTrack, but still exhibits noticeable localization bias and scale inconsistency when the target becomes extremely small or blends into the dark background, as shown in the later frames of S0302 and S0308. AutoTrack as a correlation-filter-based method, can maintain a compact response in relatively

simple low-light frames, but once the target appearance is weakened by poor illumination or corrupted by surrounding lights, response peak shift and model contamination lead to gradual drift, which is particularly evident in S0302. In comparison, SiamAPN is relatively more stable in S0307 and in the early frames of S0302 and S0308, suggesting that its proposal-based matching mechanism retains better robustness under moderate nighttime degradation; however, under severe low-light interference and strong background illumination, it still shows position offsets or switches to nearby salient regions in later frames. Overall, these nighttime cases indicate that low-light UAV tracking failures are mainly caused by the joint effects of weak appearance cues, strong artificial illumination, and tiny target size, and they further confirm that achieving robust all-weather UAV tracking still requires jointly enhancing feature discrimination, anti-distractor capability, and deployment efficiency.

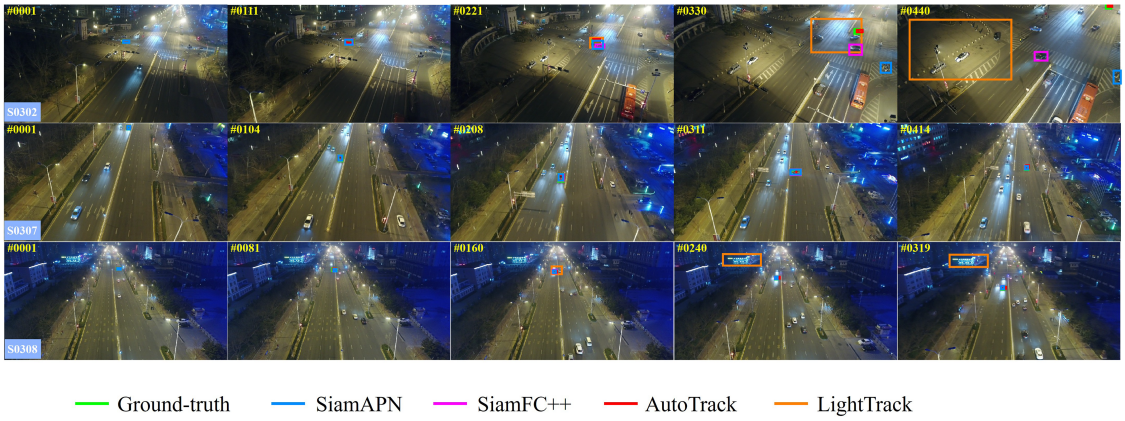

Figure 5: Comparison of typical nighttime low-light failure cases of different trackers.

The failure patterns of different trackers can be observed more explicitly across representative UAV sequences. SiamAPN generally maintains reasonable localization in relatively clear frames, but in challenging cases with tiny targets and strong distractors, such as Animal1, Horse1, and MountainBike1, it tends to suffer from target switching or gradual center deviation once nearby similar objects enter the search region. HiFT benefits from hierarchical feature interaction and usually shows stronger robustness in sequences with scale variation and motion blur, such as Surfing03 and Soccer1; however, in cluttered scenes with multiple similar instances, its predicted boxes may still become enlarged or biased, indicating that global contextual modeling can also introduce ambiguity when local target evidence is weak. TCTrack exhibits comparatively better temporal continuity under rapid motion and blur, which is reflected in Car2, Soccer1, and Surfing03; nevertheless, when the target becomes extremely small or undergoes prolonged interference from similar objects, as in Animal1 and MountainBike1, accumulated temporal errors can still lead to delayed drift. In contrast, AutoTrack, as a correlation-filter-based tracker, is more vulnerable to response peak shift and model contamination in cluttered or partially occluded scenes, so once background information is incorporated into online updating, its errors tend to persist and amplify, which is evident in Car2 and Horse1. LightTrack achieves competitive efficiency, but its lightweight design also weakens feature discrimination in highly challenging UAV scenarios; as a result, it shows the most pronounced scale overestimation and distractor capture in sequences such as Animal1, Horse1, and MountainBike1. Overall, Fig. 5 indicates that failures are mainly triggered by the combined effects of tiny targets, similar-object interference, occlusion, rapid camera motion, and motion blur, and the differences among trackers directly reflect the trade-off between representation robustness, temporal modeling, and deployment-oriented efficiency.

During evaluation, all trackers are tested using their official default parameter settings, without dataset-specific parameter tuning, to ensure fairness and reproducibility of the comparison. In addition, for fair and comprehensive comparison, Table 3 reports the Precision, AUC, and FPS of correlation filter, Siamese network, and Transformer-based trackers across different benchmarks. The top-performing methods on each dataset are highlighted using colors. It should be noted that part of the reported results are collected from the original papers, while several results are obtained through our own experimental evaluation using the corresponding benchmark toolkits.

Table 3: Performance comparison of different object tracking methods on various UAV benchmark (**red**, **green**, and **blue** for first, second, and third place, respectively).

Benchmark	UAV123			UAV123@10fps			UAVDT			UAV20L			DTB70			Methods
Tracker	P	AUC	FPS	P	AUC	FPS	P	AUC	FPS	P	AUC	FPS	P	AUC	FPS	
KCF [18]	0.523	0.331	611.7	0.406	0.265	561.1	0.571	0.290	826.2	0.311	0.196	371.2	0.468	0.280	364.1	CF
DCF [18]	0.526	0.332	861.0	0.408	0.266	811.5	0.559	0.288	1153	0.321	0.208	576.7	0.467	0.280	553.0	
SRDCF [38]	0.676	0.464	11.1	0.575	0.423	11.2	0.659	0.417	13.1	0.507	0.343	7.5	0.512	0.363	8.4	
BACF [50]	0.660	0.459	43.5	0.572	0.413	41.0	0.686	0.432	56.1	0.584	0.415	32.0	0.581	0.398	37.7	
CACF (SAMF_CA) [86]	0.605	0.415	9.3	0.523	0.365	8.7	0.559	0.303	12.0	0.537	0.352	8.7	0.532	0.345	7.1	
CACF (Staple_CA) [86]	0.672	0.454	51.6	0.587	0.420	50.1	0.695	0.394	55.5	0.497	0.345	36.9	0.504	0.351	50.7	
ECO-HC [87]	0.710	0.496	63.3	0.640	0.468	55.1	0.694	0.416	70.0	0.499	0.377	51.8	0.635	0.448	51.9	
STRCF [51]	0.681	0.481	22.6	0.627	0.457	22.4	0.629	0.411	28.6	0.575	0.410	17.4	0.649	0.437	21.9	
ARCF-H [52]	0.667	0.455	40.4	0.612	0.434	42.0	0.705	0.413	51.6	0.557	0.386	31.8	0.607	0.416	37.1	
ARCF-HC [52]	0.671	0.468	24.5	0.666	0.473	24.1	0.720	0.458	29.4	0.544	0.381	21.9	0.694	0.472	24.3	
AutoTrack [53]	0.689	0.472	48.2	0.671	0.477	47.6	0.718	0.450	56.4	0.512	0.349	44.8	0.716	0.478	48.6	
FWRDCF [34]	0.703	0.582	–	0.698	0.604	–	0.750	0.510	–	–	–	–	0.732	0.493	–	
SiamFC+_CI [88]	–	–	–	0.684	0.482	28.5	0.639	0.422	28.8	0.632	0.407	28.1	0.694	0.472	28.1	SN
SiamFC+_CX [88]	–	–	–	0.665	0.470	25.7	0.641	0.416	25.9	0.583	0.454	25.5	0.681	0.454	25.5	
SiamFC++ [89]	–	–	–	0.759	0.589	17.7	0.802	0.600	17.7	0.742	**0.575**	17.7	0.812	0.637	17.5	
DaSiamRPN [56]	0.776	0.573	–	0.692	0.483	20.6	0.794	0.498	20.4	0.631	0.442	20.6	0.705	0.474	20.4	
SiamRPN++_A [57]	–	–	–	0.737	0.551	52.0	0.774	0.557	54.4	0.701	0.533	51.5	0.793	0.586	51.6	
SiamRPN++_M [57]	–	–	–	0.771	0.578	25.7	0.757	0.556	26.2	0.723	0.547	25.4	0.785	0.593	24.7	
SiamRPN++_R [57]	–	–	–	0.784	0.594	6.2	0.801	0.594	6.3	0.758	**0.579**	6.1	0.798	0.614	6.2	
SiamGAT [90]	–	–	–	0.788	0.602	17.5	0.754	0.574	17.7	**0.796**	**0.620**	17.4	0.752	0.583	17.3	
SiamMask [91]	–	–	–	0.788	0.590	13.7	0.782	0.580	14.1	0.679	0.514	13.6	0.775	0.575	13.5	
SiamBAN [92]	–	–	–	0.770	0.585	6.2	0.806	**0.601**	6.3	0.736	0.564	6.2	0.832	0.643	6.2	
SiamCAR [79]	–	–	–	0.789	0.596	6.4	0.804	0.598	6.3	0.687	0.523	6.2	0.831	0.603	6.3	
SiamAPN [22]	0.765	0.575	–	0.760	0.571	35.3	0.710	0.516	36.2	0.721	0.539	34.5	0.784	0.586	35.0	
SiamAPN++ [23]	0.764	0.579	–	0.764	0.580	35.7	0.758	0.549	36.3	0.736	0.560	34.9	0.790	0.594	35.1	
LightTrack [58]	0.783	0.627	–	0.776	0.598	22.7	0.776	0.590	23.7	**0.791**	**0.620**	20.9	0.776	0.627	23.7	
HiFT [25]	0.787	0.589	–	0.754	0.574	–	0.652	0.475	–	**0.763**	0.566	127.0	0.802	0.594	129.9	Transformer
TCTrack [41]	0.800	0.604	–	0.774	0.588	–	0.725	0.530	–	–	–	–	0.813	0.622	125.6	
TAViT [93]	0.846	0.662	–	**0.843**	**0.660**	–	0.805	0.586	–	–	–	–	0.822	0.639	–	
SGDViT [26]	0.754	0.575	–	0.766	0.585	–	0.657	0.480	–	–	–	–	0.806	0.603	–	
Aba-ViTrack [42]	**0.864**	0.664	–	**0.850**	**0.655**	–	**0.834**	0.599	–	–	–	–	0.859	**0.664**	–	
AMST^2 [94]	0.832	0.630	–	0.798	0.616	–	–	–	–	–	–	–	0.851	**0.658**	–	
ORTrack-DeiT [32]	0.843	0.664	–	–	–	–	**0.834**	**0.601**	–	–	–	–	**0.862**	**0.664**	–	
HiT [61]	0.806	0.638	–	0.809	0.643	–	–	–	–	–	–	–	0.751	0.592	–	
LightFC [83]	**0.879**	**0.674**	–	0.813	0.637	–	–	–	–	–	–	–	**0.863**	0.652	–	
SSTrack [33]	**0.876**	**0.676**	–	**0.854**	**0.664**	–	**0.828**	**0.610**	–	–	–	–	0.855	**0.664**	–	
SGLATrack-DeiT [95]	0.849	0.669	–	–	–	–	0.819	0.599	–	–	–	–	0.844	0.651	–	
TADMT [96]	0.848	**0.682**	–	–	–	–	**0.864**	**0.699**	–	–	–	–	**0.873**	**0.697**	–	

From the perspective of method categories, CF-based trackers exhibit a clear advantage in computational efficiency, meeting the real-time requirements of UAV platforms. Consequently, they remain practically valuable in scenarios where speed is prioritized. However, the precision of these methods is relatively limited under small target, severe scale variation, and complex background conditions, as they are prone to performance degradation caused by target size changes and background interference. Some

improved CF methods introduce spatial regularization, multi-constraint optimization, and context-aware enhancements to improve accuracy, but a trade-off between speed and precision remains necessary.

In contrast, deep Siamese network-based methods demonstrate superior precision and robustness, particularly suitable for UAV scenarios involving small targets, dramatic scale changes, and complex backgrounds. For instance, the SiamAPN series, SiamRPN++ and their variants leverage deep feature representations and end-to-end matching mechanisms to achieve more accurate target localization and maintain tracking continuity, even under rapid motion and occlusion. Nonetheless, these methods generally require higher computational resources, and their real-time deployment on embedded UAV platforms necessitates additional optimization.

Recently, Transformer-based trackers have further improved performance in complex scenarios. By modeling global context and employing cross-frame attention mechanisms, Transformer methods enhance discrimination against complex background interference and multi-scale targets, and show advantages in occlusion recovery and long-sequence tracking. For example, Aba-ViTrack and TAViT achieve strong performance across multiple UAV datasets, highlighting the potential of global feature modeling for UAV target tracking. However, some large-scale Transformer models demand significant computational resources, which still poses challenges for real-time deployment on low-power UAV platforms.

Examining performance across different datasets, UAV123 and its low-frame-rate variant primarily feature clear targets and minimal occlusion, where Siamese and Transformer-based methods consistently maintain high precision and stability. In contrast, in UAVDT and UAV20L, which involve fast target motion, complex backgrounds, and frequent occlusion, the accuracy of CF-based methods decreases noticeably, whereas deep learning and Transformer methods remain robust. For DTB70, characterized by small targets and rapid scale variation, the advantages of deep Siamese networks and Transformer trackers are more pronounced, with stable and reliable performance in small target detection and multi-scale discrimination.

Fig. 6 compares the performance of several representative Siamese network trackers on UAV123@10fps, UAVDT, and DTB70. The differences in precision and success rates among these trackers mainly originate from their architectural design choices in feature representation, localization strategy, and scale handling.

SiamFC++ improves upon the original SiamFC framework by introducing deeper convolutional features and more effective feature fusion strategies. This design enhances the discriminative ability of the learned representation, allowing the tracker to better distinguish the target from background clutter. As a result, SiamFC++ typically achieves higher precision than early Siamese trackers, especially in scenarios where background interference is significant.

SiamRPN enhances localization accuracy by incorporating a region proposal network into the Siamese matching framework. The proposal mechanism enables the tracker to directly regress target bounding boxes instead of relying solely on response map peaks, which improves center localization accuracy. However, when UAV motion causes large inter-frame displacement, the fixed search region may still limit tracking robustness.

SiamBAN further improves Siamese tracking by introducing a more balanced architecture for classification and regression branches. By strengthening feature interaction between the template and search region, SiamBAN achieves more stable localization under moderate scale variation and appearance changes. This contributes to improved precision and success rates on several UAV datasets.

SiamCAR adopts an anchor-free localization mechanism, which eliminates the dependence on predefined anchor boxes. This design reduces anchor mismatch issues and allows the tracker to produce more accurate center predictions. Consequently, SiamCAR often achieves strong precision in scenarios with moderate scale variation and smooth target motion.

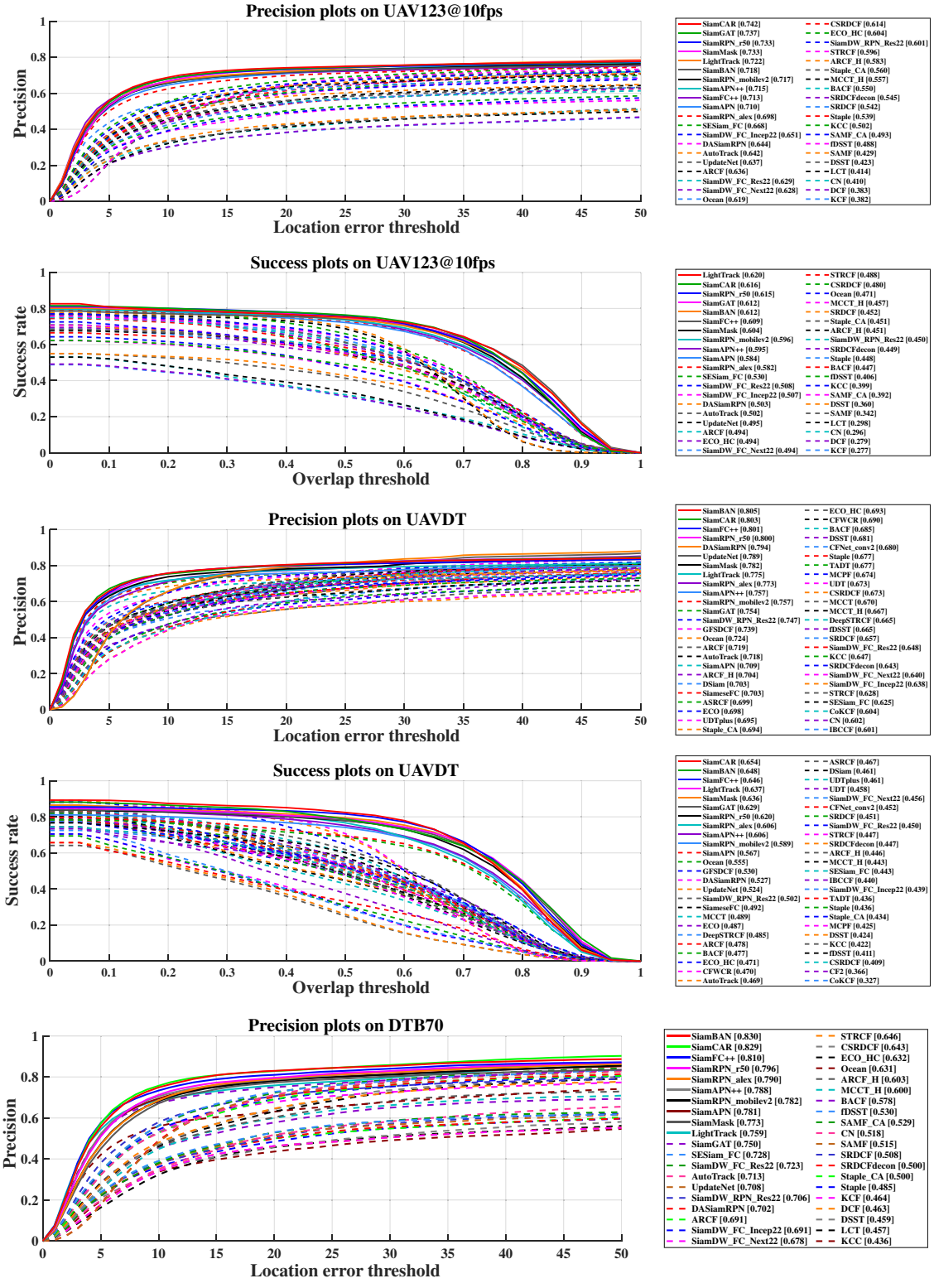

Figure 6: (Continued)

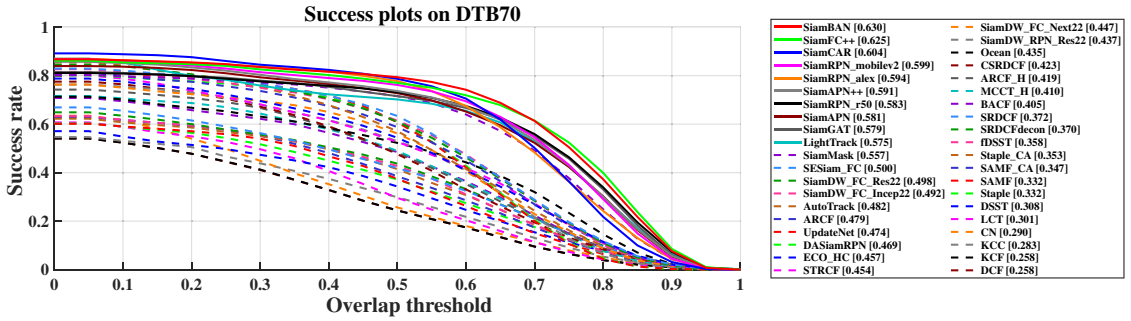

Figure 6: Overall performance of each tracker on UAV123@10fps, UAVDT and DTB70.

LightTrack focuses on improving efficiency while maintaining competitive tracking accuracy. Through neural architecture search, LightTrack obtains a lightweight network structure that balances representation capability and computational cost. This design enables the tracker to maintain stable performance while achieving higher inference speed, which is particularly important for real-time UAV applications.

Overall, different methods offer distinct advantages in UAV tracking: CF-based methods are suitable for scenarios requiring extremely high FPS and real-time performance, but their precision is limited; deep Siamese network-based trackers achieve high accuracy and robustness for small targets, scale variations, and rapid motion, albeit with relatively lower FPS; Transformer-based trackers further enhance precision in complex scenarios and occlusion recovery, but their computational efficiency still requires optimization, making lightweight and acceleration strategies critical for practical deployment.

5 Future Development Directions

With the accelerated deployment of UAV platforms in applications such as inspection, security surveillance, emergency response, and traffic monitoring, research on UAV target tracking is shifting from "improving benchmark performance" toward "enhancing practical task usability" [4,12]. At present, pursuing high precision or success rates on a single dataset is no longer sufficient to fully reflect a method's stability and generalization capability under complex aerial conditions. Future research thus requires a verifiable technical closed loop that addresses three critical aspects: onboard constraints, complex environmental conditions, and cross-scene transfer. Integrating current research trends and engineering demands, future development can be broadly summarized along three main directions: lightweight deployment, multimodal fusion, and large-model-driven paradigms.

5.1 Lightweight Deployment

UAV platforms are highly sensitive to computational resources, power consumption, and storage capacity, which limits the practical value of "high-precision but non-deployable" trackers. Future lightweight research should emphasize end-side reproducible solutions rather than merely compressing parameters or reducing FLOPs. This includes task-specific lightweight architecture search and UAV-friendly network design [58]. Meanwhile, the design of efficient attention or correlation modeling modules must consider embedded inference operator support and memory access patterns, avoiding theoretically lightweight designs that fail to run stably in practical deployment [81,82]. Given the challenges of small targets and rapid viewpoint changes in UAV scenarios, maintaining sufficient discriminative capability and robustness without significantly increasing latency will remain a key challenge for lightweight deployment.

5.2 Multimodal Fusion

In typical UAV scenarios such as low-light, backlit, smoky, or long-distance small targets, single RGB information is often insufficient to support long-term stable tracking. Introducing multimodal data, such as thermal infrared [14] or RGB-T [97], will become a critical pathway to improving robustness. Future research should go beyond "simple concatenation-based fusion" and focus on mechanisms that can adaptively adjust modality contributions according to environmental changes while preserving target identity under occlusion and interference [98]. Additionally, integrating visual prompts [99] and vision-language conditioning [100] can provide stronger semantic constraints when target appearance changes drastically or when dense distractors are present, promoting multimodal tracking from a "supplementary information" role toward "task-level redundancy".

5.3 Closed-Loop Control and Optimization

UAV target tracking in real deployments is usually a closed-loop system rather than an isolated vision module. Tracking outputs are used to command the gimbal or the UAV body, and the resulting motion changes viewpoint, motion blur, and target scale, which can reshape the five difficulty patterns discussed in this survey. Representative UAV studies have demonstrated this closed-loop coupling through vision-based tracking and landing tasks, where the visual estimation directly drives the vehicle motion and the vehicle motion in turn affects the visual observations. For example, dynamic image-based visual servo control has been used to land a quadrotor on a moving target, highlighting how control design interacts with visual feedback quality and tracking stability [101]. Similar vision-based target tracking and autonomous landing systems further indicate that stable closed-loop performance requires considering both perception reliability and flight control behavior under realistic motion and sensing conditions [102].

From a practical viewpoint, control strategies that improve stabilization can reduce apparent motion and blur and help keep the target inside the search region, which directly benefits trackers that rely on local search. On the other hand, aggressive maneuvers and weak stabilization can increase inter-frame displacement and blur, making local-search trackers more prone to loss. This motivates explicit consideration of gimbal and platform control in UAV tracking systems. In this regard, model predictive control has been studied for UAV-mounted multi-axis gimbals to achieve robust target tracking under external disturbances, which provides a representative example of using advanced control to stabilize the camera view and improve tracking feasibility [103]. Measurement delay and sampling effects are another practical factor that couples tracking and control, and robust visual servoing designs that explicitly account for visual measurement delay on inertially stabilized platforms provide useful insights for UAV gimbal or stabilized-camera settings where delay can degrade closed-loop tracking performance [104]. Beyond rotary-wing UAVs, image-based visual servo tracking control has also been investigated for fixed-wing UAVs with a monocular pan-tilt camera, showing that tracking-oriented control must respect UAV motion constraints while maintaining the target in the image plane [105].

Optimization methods complement control design at a higher decision level and can further improve system-level tracking performance. One important role of optimization is flight trajectory and viewpoint planning, where the UAV chooses motion and viewing geometry that reduce the occurrence of small targets, severe occlusion, and rapid viewpoint changes. Evolutionary computation and related meta-heuristic optimization have been widely studied for UAV path planning in complex environments, providing representative tools for selecting feasible trajectories under multiple constraints and objectives, and these results suggest clear potential to plan UAV motion in a tracking-friendly manner [106]. Another role is resource-aware deployment optimization. Although this survey mainly focuses on tracking algorithms, the UAV system must often balance tracking robustness with onboard computation and energy constraints,

and therefore it is valuable to view control, planning, and tracking as a coupled stack. A broad survey of quadrotor configurations and control has emphasized the diversity of UAV control designs and constraints, which further supports the need to incorporate control and optimization considerations when discussing UAV tracking in real applications [107]. Overall, integrating tracking, control, and optimization in a unified view strengthens the deployment-oriented understanding of UAV tracking and motivates future work on closed-loop evaluation protocols and adaptive resource-aware tracking strategies.

5.4 Large-Model-Driven Paradigms

Foundation models and large-model paradigms are reshaping the training and transfer logic of visual tasks, and UAV tracking will increasingly emphasize cross-scene generalization and low-cost adaptation [108]. Methodologically, state-of-the-art Transformer-based tracking frameworks and unified feature modeling provide a starting point for "more general tracking representations" and establish a structural basis for integrating capabilities of foundation models [30,109]. However, the computational demands of large models conflict sharply with UAV real-time deployment constraints. A more feasible direction is to combine parameter-efficient fine-tuning with lightweight inference strategies, preserving the advantages of pre-trained representations while satisfying on-board deployment requirements [110].

6 Conclusion

This paper provides a systematic review of research progress in the field of UAV visual target tracking and presents a comprehensive analysis of the applicability of different technical approaches under complex aerial scenarios, guided by the specific demands of UAV applications. By summarizing the typical challenges, it is evident that UAV target tracking differs significantly from general visual tracking tasks in terms of target scale, motion patterns, background complexity, and onboard computational constraints. These differences dictate the unique design considerations and performance trade-offs inherent to UAV tracking methods.

From the perspective of methodological evolution, traditional correlation filter (CF) approaches, owing to their computational efficiency and ease of implementation, still hold value in scenarios demanding lightweight and real-time operation. However, their performance improvement potential is limited under conditions of small targets, complex backgrounds, and prolonged occlusions. Deep learning-based Siamese network methods have achieved substantial advances in feature representation and robustness, and have become the mainstream solution for current UAV tracking applications. More recently, Transformer-based architectures, by incorporating global modeling and temporal context modeling capabilities, have demonstrated promising potential in complex UAV scenarios. Nevertheless, their computational cost and deployment challenges on edge platforms remain to be fully addressed.

Quantitative evaluations and representative failure case analyses on multiple benchmarks, including UAV123, UAV123@10fps, UAVDT, UAV20L, and DTB70, indicate that different methods still exhibit clear trade-offs among precision, stability, and real-time performance. Achieving optimal performance across all challenging conditions remains difficult, suggesting that UAV target tracking research is still in an active exploration phase and that a fully "long-term stable, all-weather, on-board deployable" solution has yet to be realized.

Future research in UAV target tracking is expected to continue along four main directions: lightweight deployment, multimodal fusion, closed-loop control and optimization, and large-model-driven paradigms. Effectively achieving long-term robust tracking in complex environments while adhering to onboard resource constraints will be key to bridging the gap between experimental evaluation and practical deployment. The systematic review and analysis presented in this paper provide a valuable reference for guiding subsequent research in this domain.

Beyond listing future directions, it is also important to discuss their scalability for practical UAV deployment. For lightweight deployment, scalability mainly means that the tracking system can adapt to UAV platforms with different payload capacities and onboard compute budgets. A practical strategy is to build a tiered model portfolio and select an appropriate model according to available resources, while keeping the same interface and training recipe. In addition, dynamic inference and hardware-aware compression can further improve scalability by adjusting computation under changing flight conditions, so that real-time performance remains stable even when onboard load or power budget varies.

For multimodal fusion, scalability depends on whether the fusion design remains compatible when new sensors are added or when some sensors are temporarily unavailable. A modular fusion pipeline is preferable, where each modality is encoded independently and fused through a lightweight interaction module. This design makes it easier to extend from RGB-only tracking to RGB plus thermal, depth, or inertial cues, and it also improves robustness under sensor dropout, misalignment, and time synchronization errors, which are common in real UAV systems.

For large-model-driven approaches, scalability is closely related to cross-scenario transfer and data efficiency. UAV tracking often suffers from limited labeled data in specific environments, and performance may degrade when viewpoint, background, or illumination shifts. Large models with stronger prior knowledge can reduce this dependency by enabling better generalization and supporting adaptation with fewer labeled samples. In practice, this can be achieved through pretraining on diverse data, self-supervised learning on unlabeled flight videos, and knowledge transfer to lightweight trackers through distillation, so that deployment remains feasible while improving robustness across scenarios.

In addition, scalability should be considered from a closed-loop system perspective that connects tracking with UAV control and optimization. In real deployment, tracking outputs drive gimbal or vehicle commands, and the resulting motion changes viewpoint, blur, and target scale, which can amplify or mitigate tracking difficulty. Therefore, scalable UAV tracking requires not only an accurate visual module, but also stable control and resource-aware optimization. Practical directions include delay-aware control and stabilization to reduce apparent motion, adaptive confidence-based policies to avoid aggressive updates under occlusion, and optimization-based planning that selects tracking-friendly viewpoints while respecting energy and safety constraints. Such closed-loop considerations are essential for transferring benchmark-level performance into robust long-term onboard operation.

To summarize the key takeaways concisely:

- UAV tracking remains strongly scenario dependent, and no single tracker family consistently dominates across small targets, motion blur, clutter, occlusion, low light, and onboard constraints.
- Correlation-filter trackers are competitive in efficiency but drift more easily under compound challenges.
- Siamese trackers provide strong general robustness but are vulnerable to large displacement with blur and low-resolution clutter.
- Transformer-based trackers show clear potential in complex backgrounds, yet real-time onboard deployment is still limited by the robustness-efficiency trade-off.
- Future work should improve stability and recoverability with deployable efficiency.

Overall, practical UAV tracking calls for designs that improve stability and recoverability under compound challenges while maintaining deployable efficiency.

Acknowledgement: None.

Funding Statement: This work was supported by the National Natural Science Foundation of China (No. 62306049, No. W2421089 and No. 92471207), the General Program of Chongqing Natural Science Foundation (No. CSTB2023NSCQ-MSX0665 and No. CSTB2023NSCQ-MSX0992), and the Fundamental Research Funds for the Central Universities (No. 2024CDJXY008).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Bo Huang and Lingyu Jin; methodology, Lingyu Jin and Rui Wang; validation, Lingyu Jin; formal analysis, Lingyu Jin; investigation, Lingyu Jin and Rui Wang; writing—original draft preparation, Lingyu Jin, Rui Wang and Bo Huang; writing—review and editing, Lingyu Jin and Bo Huang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lee KH, Hwang JN, Okopal G, Pitton J. Ground-moving-platform-based human tracking using visual SLAM and constrained multiple kernels. IEEE Trans Intell Transport Syst. 2016;17(12):3602–12. doi:10.1109/tits.2016.2557763.
2. Ababsa F, Maidi M, Didier JY, Mallem M. Vision-based tracking for mobile augmented reality. In: Multimedia services in intelligent environments. Berlin/Heidelberg, Germany: Springer; 2008. p. 297–326. doi:10.1007/978-3-540-78502-6_12.
3. Robin C, Lacroix S. Multi-robot target detection and tracking: taxonomy and survey. Auton Rob. 2016;40(4):729–60. doi:10.1007/s10514-015-9491-7.
4. Chen F, Wang X, Zhao Y, Lv S, Niu X. Visual object tracking: a survey. Comput Vis Image Underst. 2022;222:103508. doi:10.1016/j.cviu.2022.103508.
5. Luo J, Han Y, Fan L. Underwater acoustic target tracking: a review. Sensors. 2018;18(1):112. doi:10.3390/s18010112.
6. Fu C, Lu K, Zheng G, Ye J, Cao Z, Li B, et al. Siamese object tracking for unmanned aerial vehicle: a review and comprehensive analysis. Artif Intell Rev. 2023;56(1):1417–77. doi:10.1007/s10462-023-10558-5.
7. Wu X, Li W, Hong D, Tao R, Du Q. Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey. IEEE Geosci Remote Sens Mag. 2022;10(1):91–124. doi:10.1109/mgrs.2021.3115137.
8. Fu C, Li B, Ding F, Lin F, Lu G. Correlation filters for unmanned aerial vehicle-based aerial tracking: a review and experimental evaluation. IEEE Geosci Remote Sens Mag. 2022;10(1):125–60. doi:10.1109/mgrs.2021.3072992.
9. Wu P, Li Y, Xue D. UAV target tracking: a survey. Artif Intell Rev. 2025;58(11):358. doi:10.1007/s10462-025-11348-x.
10. Taufique AMN, Minnehan B, Savakis A. Benchmarking deep trackers on aerial videos. Sensors. 2020;20(2):547. doi:10.3390/s20020547.
11. Li S, Yeung DY. Visual object tracking for unmanned aerial vehicles: a benchmark and new motion models. Proc AAAI Conf Artif Intell. 2017;31(1):4140–6. doi:10.1609/aaai.v31i1.11205.
12. Zheng G, Fu C, Ye J, Li B, Lu G, Pan J. Siamese object tracking for vision-based UAM approaching with pairwise scale-channel attention. In: Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2022 Oct 23–27; Kyoto, Japan. p. 10486–92. doi:10.1109/iros47612.2022.9982189.
13. Du D, Qi Y, Yu H, Yang Y, Duan K, Li G, et al. The unmanned aerial vehicle benchmark: object detection and tracking. In: Computer vision—ECCV 2018. Cham, Switzerland: Springer International Publishing; 2018. p. 375–91. doi:10.1007/978-3-030-01249-6_23.
14. Huang B, Li J, Chen J, Wang G, Zhao J, Xu T. Anti-UAV410: a thermal infrared benchmark and customized scheme for tracking drones in the wild. IEEE Trans Pattern Anal Mach Intell. 2024;46(5):2852–65. doi:10.1109/tpami.2023.3335338.

15. Li B, Fu C, Ding F, Ye J, Lin F. All-day object tracking for unmanned aerial vehicle. IEEE Trans Mobile Comput. 2023;22(8):4515–29. doi:10.1109/tmc.2022.3162892.
16. Li X, Wang K, Wang W, Li Y. A multiple object tracking method using Kalman filter. In: Proceedings of the 2010 IEEE International Conference on Information and Automation; 2010 Jun 20–23; Harbin, China. p. 1862–6. doi:10.1109/icinfa.2010.5512258.
17. Yao H. A Survey for target tracking on Meanshift algorithms. In: Proceedings of the 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE); 2021 Jan 15–17; Guangzhou, China. p. 476–9. doi:10.1109/iccece51280.2021.9342102.
18. Henriques JF, Caseiro R, Martins P, Batista J. High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell. 2015;37(3):583–96. doi:10.1109/tpami.2014.2345390.
19. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS. Fully-convolutional Siamese networks for object tracking. In: Computer vision—ECCV 2016 workshops. Cham, Switzerland: Springer International Publishing; 2016. p. 850–65. doi:10.1007/978-3-319-48881-3_56.
20. Li B, Yan J, Wu W, Zhu Z, Hu X. High performance visual tracking with siamese region proposal network. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 8971–80. doi:10.1109/cvpr.2018.00935.
21. Voigtlaender P, Luiten J, Torr PHS, Leibe B. Siam R-CNN: visual tracking by re-detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 6578–88. doi:10.1109/cvpr42600.2020.00661.
22. Fu C, Cao Z, Li Y, Ye J, Feng C. Siamese anchor proposal network for high-speed aerial tracking. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 May 30–Jun 5; Xi'an, China. p. 510–6. doi:10.1109/icra48506.2021.9560756.
23. Cao Z, Fu C, Ye J, Li B, Li Y. SiamAPN++: siamese attentional aggregation network for real-time UAV tracking. In: Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2021 Sep 27–Oct 1; Prague, Czech Republic. p. 3086–92. doi:10.1109/iros51168.2021.9636309.
24. Danelljan M, Van Gool L, Timofte R. Probabilistic regression for visual tracking. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 7183–92. doi:10.1109/cvpr42600.2020.00721.
25. Cao Z, Fu C, Ye J, Li B, Li Y. HiFT: Hierarchical feature transformer for aerial tracking. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 15457–66. doi:10.1109/iccv48922.2021.01517.
26. Yao L, Fu C, Li S, Zheng G, Ye J. SGDViT: saliency-guided dynamic vision transformer for UAV tracking. arXiv:230304378. 2023.
27. Yan B, Peng H, Fu J, Wang D, Lu H. Learning spatio-temporal transformer for visual tracking. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 10448–57. doi:10.1109/iccv48922.2021.01028.
28. Wang N, Zhou W, Wang J, Li H. Transformer meets tracker: exploiting temporal context for robust visual tracking. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 1571–80. doi:10.1109/cvpr46437.2021.00162.
29. Cui Y, Jiang C, Wang L, Wu G. MixFormer: end-to-end tracking with iterative mixed attention. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 13608–18. doi:10.1109/cvpr52688.2022.01324.
30. Cui Y, Song T, Wu G, Wang L. Mixformerv2: efficient fully transformer tracking. Adv Neural Inf Process Syst. 2023;36:58736–51.
31. Lin L, Fan H, Zhang Z, Xu Y, Ling H. Swintrack: a simple and strong baseline for transformer tracking. Adv Neural Inf Process Syst. 2022;35:16743–54.
32. Wu Y, Wang X, Yang X, Liu M, Zeng D, Ye H, et al. Learning occlusion-robust vision transformers for real-time UAV tracking. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 17103–13. doi:10.1109/cvpr52734.2025.01594.

33. Kou Y, Lin S, Li L, Li B, Hu W, Gao J. SSTrack: sample-interval scheduling for lightweight visual object tracking. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence; 2025 Aug 16–22; Montreal, QC, Canada. p. 1314–22. doi:10.24963/ijcai.2025/147.
34. Wang X, Ma F, Wang X, Chen C. Learning feature-weighted regularization discriminative correlation filters for real-time UAV tracking. Signal Process. 2025;228(12):109765. doi:10.1016/j.sigpro.2024.109765.
35. Yuan S, Yang Y, Nguyen TH, Nguyen TM, Yang J, Liu F, et al. MMAUD: a comprehensive multi-modal anti-UAV dataset for modern miniature drone threats. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 2745–51. doi:10.1109/icra57147.2024.10610957.
36. Yue M, Zhang L, Huang J, Zhang H. Lightweight and efficient tiny-object detection based on improved YOLOv8n for UAV aerial images. Drones. 2024;8(7):276. doi:10.3390/drones8070276.
37. Bolme D, Beveridge JR, Draper BA, Lui YM. Visual object tracking using adaptive correlation filters. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010 Jun 13–18; San Francisco, CA, USA. p. 2544–50. doi:10.1109/cvpr.2010.5539960.
38. Danelljan M, Hager G, Khan FS, Felsberg M. Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 4310–8. doi:10.1109/iccv.2015.490.
39. Danelljan M, Bhat G, Khan FS, Felsberg M. ATOM: accurate tracking by overlap maximization. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 4660–9. doi:10.1109/cvpr.2019.00479.
40. Bhat G, Danelljan M, Van Gool L, Timofte R. Learning discriminative model prediction for tracking. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 6182–91. doi:10.1109/iccv.2019.00628.
41. Cao Z, Huang Z, Pan L, Zhang S, Liu Z, Fu C. TCTrack: temporal contexts for aerial tracking. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 14798–808. doi:10.1109/cvpr52688.2022.01438.
42. Li S, Yang Y, Zeng D, Wang X. Adaptive and background-aware vision transformer for real-time UAV tracking. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 13989–4000. doi:10.1109/iccv51070.2023.01286.
43. Wu Y, Li Y, Liu M, Wang X, Yang X, Ye H, et al. Learning an adaptive and view-invariant vision transformer for real-time UAV tracking. IEEE Trans Circuits Syst Video Technol. 2026;36(2):2403–18. doi:10.1109/tcsvt.2025.3599856.
44. Wei Q, Zeng B, Liu J, He L, Zeng G. LiteTrack: layer pruning with asynchronous feature extraction for lightweight and efficient visual tracking. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 4968–75. doi:10.1109/icra57147.2024.10610022.
45. Javed S, Danelljan M, Khan FS, Khan MH, Felsberg M, Matas J. Visual object tracking with discriminative filters and Siamese networks: a survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2022;45(5):6552–74. doi:10.1109/tpami.2022.3212594.
46. Marvasti-Zadeh SM, Cheng L, Ghanei-Yakhdan H, Kasaei S. Deep learning for visual tracking: a comprehensive survey. IEEE Trans Intell Transport Syst. 2022;23(5):3943–68. doi:10.1109/tits.2020.3046478.
47. Kugarajeevan J, Kokul T, Ramanan A, Fernando S. Transformers in single object tracking: an experimental survey. IEEE Access. 2023;11:80297–326. doi:10.1109/access.2023.3298440.
48. Sun N, Zhao J, Shi Q, Liu C, Liu P. Moving target tracking by unmanned aerial vehicle: a survey and taxonomy. IEEE Trans Ind Inf. 2024;20(5):7056–68. doi:10.1109/tii.2024.3363084.
49. Cao Z, Kooistra L, Wang W, Guo L, Valente J. Real-time object detection based on UAV remote sensing: a systematic literature review. Drones. 2023;7(10):620. doi:10.3390/drones7100620.
50. Galoogahi HK, Fagg A, Lucey S. Learning background-aware correlation filters for visual tracking. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 1135–43. doi:10.1109/iccv.2017.129.

51. Li F, Tian C, Zuo W, Zhang L, Yang MH. Learning spatial-temporal regularized correlation filters for visual tracking. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4904–13. doi:10.1109/cvpr.2018.00515.
52. Huang Z, Fu C, Li Y, Lin F, Lu P. Learning aberrance repressed correlation filters for real-time UAV tracking. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 2891–900. doi:10.1109/iccv.2019.00298.
53. Li Y, Fu C, Ding F, Huang Z, Lu G. AutoTrack: towards high-performance visual tracking for UAV with automatic spatio-temporal regularization. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11923–32. doi:10.1109/cvpr42600.2020.01194.
54. Lin F, Fu C, He Y, Guo F, Tang Q. BiCF: learning bidirectional incongruity-aware correlation filter for efficient UAV object tracking. In: Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31–Aug 31; Paris, France. p. 2365–71. doi:10.1109/icra40945.2020.9196530.
55. Fu C, Ye J, Xu J, He Y, Lin F. Disruptor-aware interval-based response inconsistency for correlation filters in real-time aerial tracking. IEEE Trans Geosci Remote Sens. 2021;59(8):6301–13. doi:10.1109/tgrs.2020.3030265.
56. Zhu Z, Wang Q, Li B, Wu W, Yan J, Hu W. Distractor-aware Siamese networks for visual object tracking. In: Computer VIsion—ECCV 2018. Cham, Switzerland: Springer; 2018. p. 103–19. doi:10.1007/978-3-030-01240-3_7.
57. Li B, Wu W, Wang Q, Zhang F, Xing J, Yan J. SiamRPN++: evolution of Siamese visual tracking with very deep networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 4282–91. doi:10.1109/cvpr.2019.00441.
58. Yan B, Peng H, Wu K, Wang D, Fu J, Lu H. LightTrack: finding lightweight neural networks for object tracking via one-shot architecture search. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 15180–9. doi:10.1109/cvpr46437.2021.01493.
59. Ye J, Fu C, Zheng G, Paudel DP, Chen G. Unsupervised domain adaptation for nighttime aerial tracking. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 8896–905. doi:10.1109/cvpr52688.2022.00869.
60. Lu K, Fu C, Wang Y, Zuo H, Zheng G, Pan J. Cascaded denoising transformer for UAV nighttime tracking. IEEE Robot Autom Lett. 2023;8(6):3142–9. doi:10.1109/lra.2023.3264711.
61. Kang B, Chen X, Wang D, Peng H, Lu H. Exploring lightweight hierarchical vision transformers for efficient visual tracking. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 9612–21. doi:10.1109/iccv51070.2023.00881.
62. Danelljan M, Häger G, Khan F, Felsberg M. Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference; 2014 Sep 1–5; Nottingham, UK.
63. Zeng N, Wu P, Wang Z, Li H, Liu W, Liu X. A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. IEEE Trans Instrum Meas. 2022;71:1–14. doi:10.1109/tim.2022.3153997.
64. Danelljan M, Hager G, Khan FS, Felsberg M. Discriminative scale space tracking. IEEE Trans Pattern Anal Mach Intell. 2017;39(8):1561–75. doi:10.1109/tpami.2016.2609928.
65. Xue Y, Jin G, Shen T, Tan L, Wang N, Gao J, et al. SmallTrack: wavelet pooling and graph enhanced classification for UAV small object tracking. IEEE Trans Geosci Remote Sens. 2023;61(11):1–15. doi:10.1109/tgrs.2023.3305728.
66. Williams T, Li R. Wavelet pooling for convolutional neural networks. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada.
67. Liu Q, He Z, Li X, Zheng Y. PTB-TIR: a thermal infrared pedestrian tracking benchmark. IEEE Trans Multimedia. 2020;22(3):666–75. doi:10.1109/tmm.2019.2932615.
68. Fu C, Yang X, Li F, Xu J, Liu C, Lu P. Learning consistency pursued correlation filters for real-time UAV tracking. In: Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 24–2021 Jan 24; Las Vegas, NV, USA. p. 8293–300. doi:10.1109/iros45743.2020.9340954.
69. Wu Y, Wang X, Zeng D, Ye H, Xie X, Zhao Q, et al. Learning motion blur robust vision transformers for real-time UAV tracking. Expert Syst Appl. 2026;297(2):129445. doi:10.1016/j.eswa.2025.129445.

70. Zhang P, Zhao J, Wang D, Lu H, Ruan X. Visible-thermal UAV tracking: a large-scale benchmark and new baseline. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 8886–95. doi:10.1109/cvpr52688.2022.00868.
71. Li B, Peng F, Hui T, Wei X, Wei X, Zhang L, et al. RGB-T tracking with template-bridged search interaction and target-preserved template updating. IEEE Trans Pattern Anal Mach Intell. 2025;47(1):634–49. doi:10.1109/tpami.2024.3475472.
72. Wei X, Bai Y, Zheng Y, Shi D, Gong Y. Autoregressive visual tracking. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 9697–706. doi:10.1109/cvpr52729.2023.00935.
73. Mayer C, Danelljan M, Pani Paudel D, Van Gool L. Learning target candidate association to keep track of what not to track. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 13444–54. doi:10.1109/iccv48922.2021.01319.
74. Paul M, Danelljan M, Mayer C, Van Gool L. Robust visual tracking by segmentation. In: Computer vision-ECCV 2022. Cham, Switzerland: Springer Nature; 2022. p. 571–88. doi:10.1007/978-3-031-20047-2_33.
75. Xue Y, Shen T, Jin G, Tan L, Wang N, Wang L, et al. Handling occlusion in UAV visual tracking with query-guided redetection. IEEE Trans Instrum Meas. 2024;73:1–17. doi:10.1109/tim.2024.3440378.
76. Fu C, Dong H, Ye J, Zheng G, Li S, Zhao J. HighlightNet: highlighting low-light potential features for real-time UAV tracking. In: Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2022 Oct 23–27; Kyoto, Japan. p. 12146–53. doi:10.1109/iros47612.2022.9981070.
77. Zhu J, Tang H, Cheng ZQ, He JY, Luo B, Qiu S, et al. DCPT: darkness clue-prompted tracking in nighttime UAVs. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 7381–8. doi:10.1109/icra57147.2024.10610544.
78. Wu Y, Yang X, Wang X, Ye H, Zeng D, Li S. MambaNUT: n. Nighttime UAV tracking via mamba and adaptive curriculum learning. arXiv:2412.00626v3. 2024.
79. Guo D, Wang J, Cui Y, Wang Z, Chen S. SiamCAR: siamese fully convolutional classification and regression for visual tracking. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 6269–77. doi:10.1109/cvpr42600.2020.00630.
80. Zheng J, Liang M, Huang S, Ning J. Exploring the feature extraction and relation modeling for light-weight transformer tracking. In: Computer vision-ECCV 2024. Cham, Switzerland: Springer Nature; 2024. p. 110–26. doi:10.1007/978-3-031-73397-0_7.
81. Chen X, Kang B, Wang D, Li D, Lu H. Efficient visual tracking via hierarchical cross-attention transformer. In: Computer vision—ECCV 2022 workshops. Cham, Switzerland: Springer Nature; 2023. p. 461–77. doi:10.1007/978-3-031-25085-9_26.
82. Blatter P, Kanakis M, Danelljan M, Van Gool L. Efficient visual tracking with exemplar transformers. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 1571–81. doi:10.1109/wacv56688.2023.00162.
83. Li Y, Wang B, Wu X, Liu Z, Li Y. Lightweight full-convolutional Siamese tracker. Knowl Based Syst. 2024;286:111439. doi:10.1016/j.knosys.2024.111439.
84. Liu X, Xu W, Wang Q, Zhang M. Energy-efficient computing acceleration of unmanned aerial vehicles based on a CPU/FPGA/NPU heterogeneous system. IEEE Internet Things J. 2024;11(16):27126–38. doi:10.1109/jiot.2024.3397649.
85. Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking. In: Computer Vision—ECCV 2016. Cham, Switzerland: Springer International Publishing; 2016. p. 445–61. doi:10.1007/978-3-319-46448-0_27.
86. Mueller M, Smith N, Ghanem B. Context-aware correlation filter tracking. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 July 21–26; Honolulu, HI, USA. p. 1396–404. doi:10.1109/cvpr.2017.152.
87. Danelljan M, Bhat G, Khan FS, Felsberg M. ECO: efficient convolution operators for tracking. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 6638–46. doi:10.1109/cvpr.2017.733.

88. Zhang Z, Peng H. Deeper and wider Siamese networks for real-time visual tracking. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 4591–600. doi:10.1109/cvpr.2019.00472.
89. Xu Y, Wang Z, Li Z, Yuan Y, Yu G. SiamFC++: towards robust and accurate visual tracking with target estimation guidelines. Proc AAAI Conf Artif Intell. 2020;34(7):12549–56. doi:10.1609/aaai.v34i07.6944.
90. Guo D, Shao Y, Cui Y, Wang Z, Zhang L, Shen C. Graph attention tracking. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 9543–52. doi:10.1109/cvpr46437.2021.00942.
91. Wang Q, Zhang L, Bertinetto L, Hu W, Torr PHS. Fast online object tracking and segmentation: a unifying approach. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, CA, USA. p. 1328–38. doi:10.1109/cvpr.2019.00142.
92. Chen Z, Zhong B, Li G, Zhang S, Ji R. Siamese box adaptive network for visual tracking. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 6668–77. doi:10.1109/cvpr42600.2020.00670.
93. Li S, Yang X, Wang X, Zeng D, Ye H, Zhao Q. Learning target-aware vision transformers for real-time UAV tracking. IEEE Trans Geosci Remote Sens. 2024;62:1–18. doi:10.1109/tgrs.2024.3417400.
94. Park H, Lee I, Jeong D, Paik J. AMST2: aggregated multi-level spatial and temporal context-based transformer for robust aerial tracking. Sci Rep. 2023;13(1):9062. doi:10.1038/s41598-023-36131-2.
95. Xue C, Zhong B, Liang Q, Zheng Y, Li N, Xue Y, et al. Similarity-guided layer-adaptive vision transformer for UAV tracking. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 6730–40. doi:10.1109/cvpr52734.2025.00631.
96. Du G, Zhou P, Yadikar N, Aysa A, Ubul K. Mamba meets tracker: exploiting token aggregation and diffusion for robust unmanned aerial vehicles tracking. Complex Intell Syst. 2025;11(4):204. doi:10.1007/s40747-025-01821-z.
97. Xia J, Shi D, Song K, Song L, Wang X, Jin S, et al. Unified single-stage transformer network for efficient RGB-T tracking. arXiv:230813764. 2023.
98. Hui T, Xun Z, Peng F, Huang J, Wei X, Wei X, et al. Bridging search region interaction with template for RGB-T tracking. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 13630–9. doi:10.1109/cvpr52729.2023.01310.
99. Zhu J, Lai S, Chen X, Wang D, Lu H. Visual prompt multi-modal tracking. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 9516–26. doi:10.1109/cvpr52729.2023.00918.
100. Li H, Liu X, Li G. A benchmark for UAV-view natural language-guided tracking. Electronics. 2024;13(9):1706. doi:10.3390/electronics13091706.
101. Serra P, Cunha R, Hamel T, Cabecinhas D, Silvestre C. Landing of a quadrotor on a moving target using dynamic image-based visual servo control. IEEE Trans Robot. 2016;32(6):1524–35. doi:10.1109/tro.2016.2604495.
102. Hoang T, Bayasgalan E, Wang Z, Tsechpenakis G, Panagou D. Vision-based target tracking and autonomous landing of a quadrotor on a ground vehicle. In: Proceedings of the 2017 American Control Conference (ACC); 2017 May 24–26; Seattle, WA, USA. p. 5580–5. doi:10.23919/acc.2017.7963823.
103. Altan A, Hacıoğlu R. Model predictive control of three-axis gimbal system mounted on UAV for real-time target tracking under external disturbances. Mech Syst Signal Process. 2020;138(1):106548. doi:10.1016/j.ymssp.2019.106548.
104. Yang J, Liu X, Sun J, Li S. Sampled-data robust visual servoing control for moving target tracking of an inertially stabilized platform with a measurement delay. Automatica. 2022;137(1):110105. doi:10.1016/j.automatica.2021.110105.
105. Yang L, Liu Z, Wang X, Yu X, Wang G, Shen L. Image-based visual servo tracking control of a ground moving target for a fixed-wing unmanned aerial vehicle. J Intell Rob Syst. 2021;102(4):81. doi:10.1007/s10846-021-01425-y.
106. Jiang Y, Xu XX, Zheng MY, Zhan ZH. Evolutionary computation for unmanned aerial vehicle path planning: a survey. Artif Intell Rev. 2024;57(10):267. doi:10.1007/s10462-024-10913-0.

107. Shraim H, Awada A, Youness R. A survey on quadrotors: configurations, modeling and identification, control, collision avoidance, fault diagnosis and tolerant control. IEEE Aerosp Electron Syst Mag. 2018;33(7):14–33. doi:10.1109/maes.2018.160246.
108. Hong L, Yan S, Zhang R, Li W, Zhou X, Guo P, et al. OneTracker: unifying visual object tracking with foundation models and efficient tuning. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 19079–91. doi:10.1109/cvpr52733.2024.01805.
109. Ye B, Chang H, Ma B, Shan S, Chen X. Joint feature learning and relation modeling for tracking: a one-stream framework. In: Computer vision—ECCV 2022. Cham, Switzerland: Springer Nature; 2022. p. 341–57. doi:10.1007/978-3-031-20047-2_20.
110. Chen X, Peng H, Wang D, Lu H, Hu H. SeqTrack: sequence to sequence learning for visual object tracking. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 14572–81. doi:10.1109/cvpr52729.2023.01400.