Highlights

UAV-Based Maritime Object Tracking: A New Data Set and Dynamic Multi-scale Fusion Transformer

Xinyue Wang, Jingjing Xiao, Xinghui Dong

- A new UAV-acquired maritime data set, i.e., UAVSeaShip102, is collected to evaluate long-term tracking robustness under challenging sea-surface environments involving reflections, illumination changes, and unstructured backgrounds.
- Proposing a Dynamic Multi-scale Fusion Transformer Single Object Tracker, namely, DMFTrack.
- A Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM) is introduced to adaptively aggregate multi-scale features using a learnable gating mechanism, effectively handling drastic scale variations in UAV scenes.
- An Adaptive Template Update Mechanism (ATUM) is designed based on a dual-stage score head, enabling stability-aware template updates and mitigating drift during long-term aerial tracking.

UAV-Based Maritime Object Tracking: A New Data Set and Dynamic Multi-scale Fusion Transformer⋆

Xinyue Wang[a], Jingjing Xiao[b,*] and Xinghui Dong[a,*]

[a]State Key Laboratory of Physical Oceanography, Qingdao Key Laboratory of Smart Ocean, and Faculty of Information Science and Engineering, Ocean University of China, No.238 Songling Road, Qingdao, 266100, Shangdong, China
[b]Bio-Med Informatics Research Centre, The Second Affiliated Hospital of the Army Medical University, 400037, Chongqing, China

ARTICLE INFO

Keywords:
UAV-Based Maritime Tracking
Ship Tracking
Single Object Tracking
Visual Transformer
Multi-scale Feature Fusion

ABSTRACT

To perform the Unmanned Aerial Vehicle-based (UAV-based) maritime object tracking task, we acquire a new ship tracking data set, namely, UAVSeaShip102. This task faces severe challenges due to rapid target motion, drastic scale variations, and frequent occlusion in complex aerial scenes. Recent transformer-based trackers have shown remarkable potential in modeling long-range dependencies. However, their performance often deteriorates when faced with large-scale changes or unreliable updates. To address these issues, we propose a Dynamic Multi-scale Fusion Transformer Tracker (DMFTrack) for UAV-based single object tracking, which comprises two key components, i.e., a Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM) and an Adaptive Template Update Mechanism (ATUM). The DWMS-FFM adaptively aggregates features from multiple scales via a learnable gating strategy, allowing the model to dynamically balance spatial details and global context. Furthermore, the ATUM employs a Score Head to estimate the reliability of the current template through dual-stage cross-attention, allowing confidence-guided online template updates that mitigate model drift. Extensive experiments conducted on UAVSeaShip102 and five commonly-used UAV-based benchmarks demonstrate that our method outperforms 22 trackers in challenging scenarios, which involve scale variation, occlusion, and background clutter, while maintaining real-time performance. It is suggested that DMFTrack is also applicable to other aerial tracking scenarios.

1. Introduction

Unmanned vehicles have been widely applied to different maritime scenarios, such as maritime visual perception (Yao, Huang, Shi, Han and Yu, 2026), cooperative multiple Unmanned Surface Vehicle (USV) decision-making (Yu, Wang, Sun, Guo and Liang, 2026; Chen, Liang, Zhang, Liu, Yu and Li, 2025), and Unmanned Aerial Vehicle-based (UAV-based) object tracking (Wen, Meriaux, Guzmán, Wang, Shi and Dudek, 2026). UAV-based object tracking plays a crucial role in a wide range of real-world applications, such as maritime surveillance (Bai, Zhang and Xu, 2025; Cheng, Zhu and Wu, 2023), search and rescue (Li, Zhang, Jiang and Zhang, 2023; Zhao, Shen, Liang, Wang and Han, 2025), and state estimation (Novák, Báča, Procházka and Saska, 2025). Compared with ground-based tracking systems, UAV platforms provide a significantly larger field of view, flexible motion patterns, and the ability to cover wide-area environments (Kumar and Mondal, 2021). However, these advantages also introduce a series of unique challenges, including drastic appearance variations, agile camera motion, severe occlusions, strong illumination changes, and significant scale variation (Wang, Wang, Zhang and Liu, 2025).

The combination of these factors makes UAV-based Single Object Tracking (SOT) more unstable and struggling, particularly in long-term aerial scenarios where targets undergo continuous motion and environmental conditions change rapidly (Wu, Wang, Yang, Liu, Zeng, Ye and Li, 2025b). Therefore, designing a tracking method that can adaptively capture multi-scale information and maintain a stable target representation across long temporal spans remains an open but essential research problem for UAV-based SOT. In addition, rare UAV-based maritime object tracking data sets are publicly available. This issue limits the study of maritime object tracking approaches.

⋆This study was supported by Shandong Provincial Natural Science Foundation (No. ZR2025ZD36).
⋆⋆The data set, source code and model will be made publicly available on the acceptance of the paper.
*Corresponding authors
wxy6329@stu.ouc.edu.cn (X. Wang); shine636363@sina.com (J. Xiao); xinghui.dong@ouc.edu.cn (X. Dong)
ORCID(s): 0000-0001-5020-638X (J. Xiao); 0000-0003-1047-427X (X. Dong)
[1]

In recent years, Transformer-based trackers have achieved impressive progress due to their strong ability to model long-range dependencies and contextual relationships (Liu, Xiao and Dong, 2025a). The self-attention mechanism in Vision Transformers (ViTs) (Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, Dehghani, Minderer, Heigold, Gelly, Uszkoreit and Houlsby, 2021) enables effective interaction between template and search regions, producing expressive and discriminative target features (Hu, Zhao, Hui, Li and You, 2023). Despite these strengths, current Transformer-based UAV trackers still suffer from two critical limitations.

First, many existing methods rely on fixed-scale backbone features or apply simple linear fusion (Li, Yang, Wang, Zeng, Ye and Zhao, 2024a), which limits their ability to handle the massive scale variations commonly encountered in UAV-acquired videos, in particular, under significant height changes, viewpoint transitions, or rapid target motion. Second, the template in conventional trackers is either fixed or updated at a fixed interval, which may cause template drift when the tracker encounters occlusion, fast motion, or background interference (Du, Zhou, Yadikar, Aysa and Ubul, 2025). Consequently, maintaining a reliable template over long sequences remains a challenging issue.

To address the above-mentioned challenges, we first introduce UAVSeaShip102, a newly collected UAV-based maritime object tracking data set that consists of real-world sequences of ships captured using an off-the-shelf UAV over open sea environments. This data set serves as a specific maritime UAV-based tracking benchmark and provides a valuable complement to existing UAV-based tracking data sets, which are predominantly focused on urban or terrestrial scenes. Although the UAVSeaShip102 mainly contains sequences captured under clear weather, it provides a valuable foundation for studying long-term UAV-based maritime object tracking, characterized by continuous motion, dynamic backgrounds, and illumination fluctuations over water surfaces.

Then we propose a Dynamic Multi-scale Fusion Transformer Tracker (DMFTrack) that explicitly considers two complementary sources of uncertainty in UAV-based long-term tracking, including spatial-scale uncertainty and temporal template uncertainty. Rather than treating the Transformer backbone itself as the main methodological contribution, we employ a compact Transformer as a general feature extractor and focus on two task-oriented mechanisms. First, the Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM) dynamically reweights representations generated at different spatial scales according to the current feature context, enabling the tracker to balance fine-grained localization cues and broader contextual information. Second, the Adaptive Template Update Mechanism (ATUM) estimates the reliability of the current template and performs confidence-guided temporal adaptation, unlike conventional static templates (Cen and Jung, 2018) or heavy online learning strategies (Bhat, Danelljan, Gool and Timofte, 2019), thereby reducing the risk of template contamination and long-term drift. The two mechanisms address complementary spatial and temporal failure modes and are designed to be applicable beyond maritime scenes.

Although this work is primarily motivated by the challenges that UAV-based maritime object tracking faces, including severe scale variations, background interference, and long-term drift, the proposed DWMS-FFM and ATUM are not limited to maritime environments. These challenges are also commonly encountered by general UAV-based tracking tasks. In this context, the proposed framework likely exhibits a broader applicability across various aerial tracking scenarios.

The main contributions of this work can be summarized as fourfold.

- A new UAV-Based Maritime Ship Single Object Tracking Data Set (UAVSeaShip102) is acquired, which provides real-world UAV footage of ships for the study of long-duration, low-altitude tracking tasks.
- We adopt a Dynamic Multi-scale Fusion Transformer Tracker (DMFTrack), which is developed to jointly address spatial-scale variation and temporal template degradation for robust UAV-based single-target tracking in various scenarios, particularly in maritime environments.
- A lightweight Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM) is introduced. This module addresses spatial uncertainty by dynamically weighting scale-specific representations.
- We design an Adaptive Template Update Mechanism (ATUM) that addresses temporal uncertainty by estimating template reliability and performing confidence-guided template adaptation.

The rest of this paper is organized as follows. Section 2 reviews the related work to this study. Section 3 elaborates on the details of the UAVSeaShip102. The details of our proposed approach are provided in Section 4. Section 5 presents experimental setup and results. We conclude the paper in Section 6.

2. Related Work

2.1. Visual Object Tracking

Visual Object Tracking (VOT) is a fundamental problem in computer vision, aiming to estimate the state of a target object across consecutive video frames given only its initial position. Early approaches, such as Correlation Filter (CF)-based trackers (e.g., KCF (Henriques, Caseiro, Martins and Batista, 2015), ECO (Danelljan, Bhat, Khan and Felsberg, 2017a), and BACF (Galoogahi, Fagg and Lucey, 2017)), modeled target appearance with handcrafted features and achieved efficient performance but limited robustness under complex conditions. With the advent of deep learning, Siamese network-based trackers (e.g., SiamFC (Cen and Jung, 2018), SiamRPN++ (Li, Wu, Wang, Zhang, Xing and Yan, 2019), and Ocean (Zhang, Peng, Fu, Li and Hu, 2020)) have dominated the field by learning a general similarity function between the template and search regions. These methods greatly improved robustness and accuracy. However, most of those methods rely on fixed receptive fields and fail to model long-range dependencies, which limits their adaptability in dynamic UAV-based tracking environments characterized by large appearance and scale variations.

2.2. Transformer-Based Visual Object Tracking

In recent years, Transformers have emerged as powerful alternatives to convolutional architectures due to their ability to capture long-range dependencies through self-attention. Vision Transformers (ViTs) (Dosovitskiy et al., 2021) and their variants, such as ViTDet (Li, Mao, Girshick and He, 2022b), Swin-Transformer (Liu, Lin, Cao, Hu, Wei, Zhang, Lin and Guo, 2021), and DeiT (Touvron, Cord, Douze, Massa, Sablayrolles and Jegou, 2021), have been successfully adapted for tracking tasks. For example, TransT (Chen, Yan, Zhu, Wang, Yang and Lu, 2021) introduced a Transformer encoder-decoder structure to explicitly model the relation between template and search features, while STARK (Yan, Peng, Fu, Wang and Lu, 2021a) and MixFormer (Cui, Jiang, Wang and Wu, 2022) unified tracking as a sequence prediction problem within a pure Transformer framework.

More recent trackers, such as OSTrack (Ye, Chang, Ma, Shan and Chen, 2022) and SimTrack (Luo, Yang and Yuille, 2021), further simplified the pipeline by means of joint feature embedding and similarity-guided attention. Despite these advances, directly applying standard Transformers to UAV-based tracking remains challenging. Many lightweight models for UAV-based single-target tracking have also emerged, such as AVTrack (Li, Liu, Wu, Wang, Yang and Li, 2024b) and ORtrack (Wu et al., 2025b). SGLATrack (Xue, Zhong, Liang, Zheng, Li, Xue and Song, 2025) further explored adaptive layer selection to reduce redundant computation while preserving informative representations.

Beyond architectural efficiency, recent studies have investigated temporal learning and complementary visual cues. Low-cost temporal learning methods (Xue, Liang, Zhong, Zu, Xue, Xia and Song, 2026) have been developed to exploit temporal information with limited additional computational overhead, while shared-query approaches have been used to jointly model motion and appearance cues for visual tracking (Xue, Zhong, Liang, Xia and Song, 2024). To tackle challenging illumination conditions, parameter- and computation-efficient transfer learning has also been explored for nighttime UAV-based tracking (Zu, Liang, Xue, Tan, Song and Zhong, 2026). Within related aerial and satellite tracking scenarios, motion-aware correlation filtering methods (Lin, Zheng, Xue, Fu, Li and Shen, 2024) further demonstrated the importance of explicitly modeling target motion under large viewpoint and scale variations.

However, UAV videos often exhibit drastic scale variations and motion blurring, while these Transformer models extract features at a single scale and cannot adaptively emphasize scale-relevant information. In addition, those models usually lack a lightweight template update mechanism, which poses challenges in the context of rapidly moving UAVs. On the other hand, fixed or periodic template update mechanisms may cause drift when the tracker fails to assess the reliability of the current template.

2.3. Multi-scale Feature Fusion

Multi-scale feature fusion plays an essential role in visual tracking, particularly in UAV-based applications where the target size changes drastically due to aerial motion and camera zooming. Traditional methods, such as FPN (Lin, Dollár, Girshick, He, Hariharan and Belongie, 2017) and PANet (Liu, Qi, Qin, Shi and Jia, 2018), fuse hierarchical CNN features by top-down and lateral connections, while recent Transformer-based trackers attempt to combine multi-layer tokens or attention maps to integrate contextual cues. However, most of these approaches rely on static fusion strategies, where the contribution of each scale is fixed regardless of the target dynamics, or rely on heavy computation, which is not suitable for UAV platforms. In contrast, our DWMS-FFM employs a dynamic gating mechanism that generates content-aware fusion weights conditioned on the input features. By adaptively reweighting scale-specific representations, DWMS-FFM effectively captures both fine-grained details for localization and coarse-scale semantics for robust context understanding, thereby enhancing UAV-based tracking stability in challenging environments.

2.4. Template Update Mechanisms

Maintaining an accurate and up-to-date template is crucial for robust long-term tracking. Early Siamese trackers typically adopted a fixed template from the first frame, which simplifies inference but leads to failure once the target appearance changes. Later methods introduced adaptive template update schemes based on heuristic strategies or exponential moving averages (e.g., ATOM (Danelljan, Bhat, Khan and Felsberg, 2019), DiMP (Bhat et al., 2019), STARK (Yan et al., 2021a)), but these approaches lack an explicit reliability estimation and are prone to template drift when the tracker is uncertain. To mitigate this problem, some methods explored confidence-guided updates or dual-template frameworks (Liu, Zhou, Dong, Xiao and Dong, 2025b). Nevertheless, these methods usually rely on simple classification confidence as a proxy, which does not accurately reflect the similarity between template and search features. Our ATUM introduces a dedicated Score Head, which explicitly predicts the template reliability score through dual-stage cross-attention between the search and template features. The confidence-guided interpolation ensures that the template is updated only when robustness can be reliably, effectively maintained under occlusion and distractor interference, which is a critical requirement for long-term UAV-based tracking.

2.5. UAV-Based Tracking Benchmarks

Several publicly available data sets have been established for UAV-based tracking, such as UAV123 (Mueller, Smith and Ghanem, 2016), DTB70 (Li and Yeung, 2017), and UAVTrack112 (Fu, Cao, Li, Ye and Feng, 2021a), which have greatly facilitated aerial tracking research. However, these data sets primarily focus on urban and terrestrial scenes, with limited coverage of maritime or long-term sequences. SeaDronesSee (Varga, Kiefer, Messmer and Zell, 2022) comprises the data from the perspective of maritime drones. However, its main focus is on maritime search and rescue scenarios rather than the long-term tracking of various ships at sea.

To further explore UAV-based tracking robustness in maritime environments, we collected a UAV-Acquired Maritime Ship Single Object Tracking Data Set, referred to as UAVSeaShip102, which consists of long-term real-world sequences of ships captured from a low-altitude UAV. Our UAVSeaShip102 collection focuses on the scenario of using drones for long-term tracking of maritime vessels, providing a real drone perspective of the long-term single object tracking sequence of maritime vessels, and covering several representative real-world weather and maritime conditions. Although UAVSeaShip102 is not perfect due to real-world environmental limitations, it provides a valuable supplement to existing UAV-based tracking benchmarks, facilitating the evaluation of long-term robustness for maritime targets.

3. UAVSeaShip102 Data Set

To complement existing UAV-based tracking benchmarks and provide a more comprehensive evaluation environment for long-term aerial tracking, we introduce UAVSeaShip102, a maritime-focused data set specifically designed to capture the unique challenges of UAV-based ship tracking. The following subsections describe the motivation, construction process, statistical characteristics, and evaluation protocols of UAVSeaShip102 in detail.

3.1. Motivation and Overview

Although many UAV-based tracking data sets have been introduced in recent years, such as UAV123 (Mueller et al., 2016), DTB70 (Li and Yeung, 2017), and VisDrone2019-SOT (Du, Zhu, Wen, Bian, Ling, Hu, Zheng, Peng, Wang, Zhang, Bo, Shi, Zhu, Han, Zhang, Liu, Wu, Wen, Wang, Fan, Chen, Gao, Zhang, Zhou, Zhou, Wang, Wan, Kittler, Zhang, Huang, Yang, Zhang, Huang, Zhou, Shi, Ding, Wang, Wang, Hu, Laganiere, Ma, Zhang, Zou, Zhao, Li, Zhu, Li, Ge, Xuan and Xu, 2019), most of them primarily focus on land-based or urban scenes, where the background is relatively structured, and the target motion is constrained. In contrast, UAV-based maritime object tracking presents distinct challenges, such as unstructured backgrounds and long-term continuous target motion. These conditions often cause model drift and unstable template updates, making it difficult to evaluate the long-term robustness of UAV trackers.

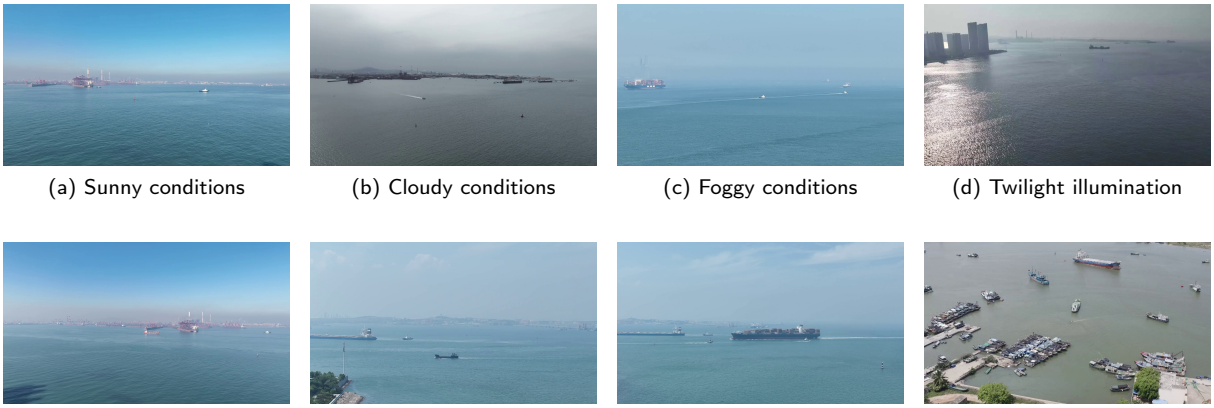
To better analyze the generalizability and stability of our proposed DMFTrack in real-world maritime conditions, we constructed a dedicated UAV-acquired data set, namely, UAVSeaShip102. This data set focused on long-term tracking of marine vessels captured from UAV platforms under various environmental conditions.

3.2. Data Collection and Annotation

The UAVSeaShip102 data set was collected using a DJI Mavic3 Classic with 102 sequences, flying at altitudes ranging from 30 to 85 meters, the complete parameters used by the UAV for video acquisition are detailed in Table 1.

Parameter	Values
ISO	100
Shutter	1/1250.0, 1/1600.0
Focal length	24, 46.8, 69.7mm
Aperture	2.8
Relative altitude	30-85m
Absolute altitude	110-165m
Resolution	1280×720
Frame rates	30 FPS

Table 1
Details of the camera used for capturing all the video sequences in the UAVSeaShip102 data set.



(a) Sunny conditions (b) Cloudy conditions (c) Foggy conditions (d) Twilight illumination

(e) Daytime illumination (f) Cruising mode (g) Turning mode (h) Docking mode

Figure 1: Eight sequence examples contained in UAVSeaShip102, captured under different weather conditions, illuminations and motion patterns.

The videos were captured under diverse environmental conditions, including sunny, slightly cloudy and foggy weather, as well as daytime and twilight illumination. Each video sequence recorded one primary target vessel, which ranged from small fishing boats to large cargo ships, performing different motion patterns, such as linear cruising, turning and docking. Eight representative examples are presented in Figure 1.

All videos undergo frame-by-frame manual annotation using an annotation tool named DarkLabel2.4. To ensure annotation accuracy, three measures were adopted. First, each sequence was annotated by personnel with professional knowledge. Second, a senior annotator performed a final quality check. Third, bounding boxes followed axis-aligned rectangle format.

3.3. Characteristics and Statistics of the Data Set

The UAVSeaShip102 contains a total of 102 sequences with over 282,000 frames, covering diverse maritime environments, such as open sea, harbor, and nearshore regions. The distribution of the number of frames computed across all the sequences can be seen in Figure 2. The target vessels in the data set exhibit substantial appearance changes, scale variations, and long-term trajectories (the length of some sequences exceeds 3,000 frames). In total, nine representative scenario attributes were recognized. The detailed description of each attribute is presented in Table 2. To facilitate challenge-specific analysis, each sequence was annotated with these attributes. Figure 3 presents the distribution of scenario attributes calculated across all the sequences.

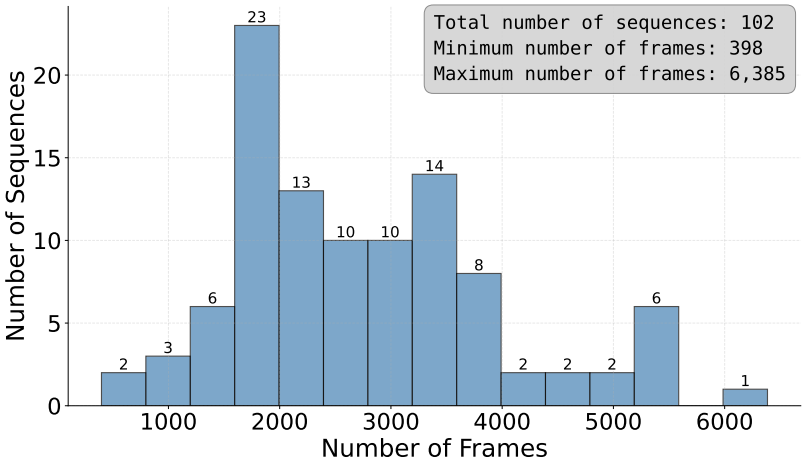


Figure 2: The distribution of number of frames computed across the sequences in the UAVSeaShip102.

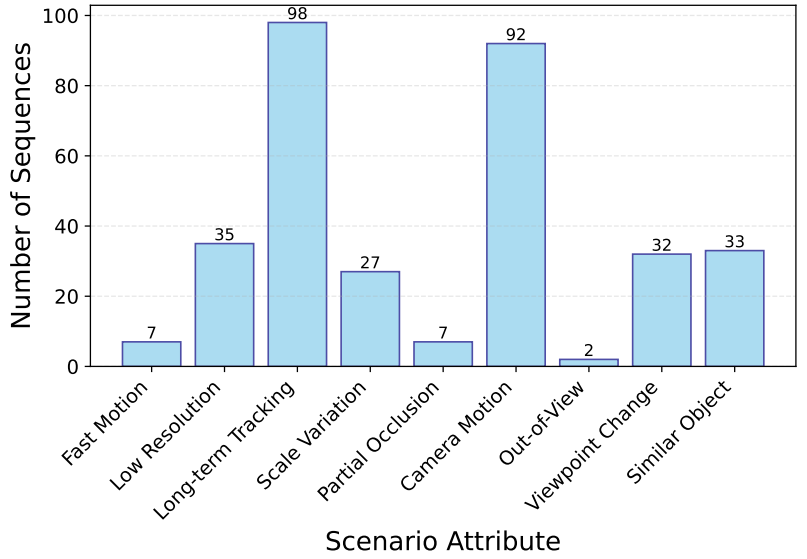


Figure 3: The distribution of scenario attributes which appear in our UAVSeaShip102.

To provide a direct comparison with established UAV-based tracking benchmarks, Table 3 summarizes the basic statistics and representative scenario characteristics of UAVSeaShip102, UAV123 (Mueller et al., 2016), DTB70 (Li and Yeung, 2017), and UAVTrack112 (Fu et al., 2021a). Specifically, UAV123 contains 123 sequences and 112,578 frames, while DTB70 contains 70 sequences and approximately 15,777 frames. UAVTrack112 contains 112 sequences and more than 100,000 frames. In contrast, UAVSeaShip102 contains 102 sequences but more than 282,000 frames, resulting in an average sequence length of over 2,765 frames. This characteristic indicates that the proposed data set places a greater emphasis on long-duration tracking than the general-purpose UAV-based tracking benchmarks tested in this study.

Through preliminary testing, we observed that the UAVSeaShip102 particularly stresses the ability of a tracker to adapt template features gracefully during long-term tracking and fuse multi-scale information robustly when the size of targets fluctuates drastically. Such characteristics introduce unique challenges for UAV-based maritime tracking. In particular, the constantly changing sea surface makes it difficult for trackers to maintain stable appearance representations. In addition, the combination of long-term trajectories, low-texture regions, and severe scale fluctuations increases the

Scenario Attribute	Description
Fast Motion	Motion of the ground-truth bounding box is of large magnitude between two consecutive frames.
Low Resolution	The ground-truth bounding box is small.
Long-Term Tracking	The sequence contains more than 1000 frames.
Scale Variation	The scale of the target varies greatly.
Partial Occlusion	The target is partially occluded.
Camera Motion	Abrupt motion of the camera.
Out-of-View	Some portion of the target leaves the view.
Viewpoint Change	Viewpoint affects target appearance significantly.
Similar Object	There is at least an object with the similar appearance near the target.

Table 2
Detailed descriptions of nine scenario attributes.

Data Set	Sequences	Total Frames	Avg. Frames/Seq.	Maritime Environment	Long-Term Tracking
UAV123 (Mueller et al., 2016)	123	112,578	915	Limited	Limited
DTB70 (Li and Yeung, 2017)	70	15,777	225	Limited	Limited
UAVTrack112 (Fu et al., 2021a)	112	>100K	>893	Limited	Extensive
UAVSeaShip102	**102**	**>282K**	**>2,765**	**Extensive**	**Extensive**

Table 3
Comparison between UAVSeaShip102 and three UAV-based single-object tracking data sets.

risk of target drift, especially when the vessel becomes visually similar to the surrounding water patterns. As a result, the UAVSeaShip102 can be considered a challenging benchmark for evaluating the robustness and adaptability of UAV-based tracking algorithms.

3.4. Discussion and Limitations

UAVSeaShip102 follows standard UAV-based tracking evaluation criteria, including Success Rate, AUC, Precision computed based on center location error and Normalized Precision (P_{Norm}) calculated for cross-resolution comparison. To maintain consistency with prior UAV-acquired benchmarks, we adopted one-shot tracking initialization, where only the first-frame annotation was used and either reinitialization or manual correction was not allowed. Trackers must manage target reappearance, occlusion recovery, and scale changes independently.

Despite its usefulness, UAVSeaShip102 still encounters three limitations. To be specific, (1) the total number of sequences is smaller than that of large-scale benchmarks, such as LaSOT (Fan, Lin, Yang, Chu, Deng, Yu, Bai, Xu, Liao and Ling, 2019) and GOT-10k (Huang, Zhao and Huang, 2021), (2) sequences captured in extreme weather conditions, such as heavy rain and severe storms, and nighttime are underrepresented, and (3) the data set focuses on maritime single object tracking and does not include multi-target interactions. However, these limitations mainly result from the practical constraints of UAV-based maritime data collection and annotation. In this context, UAVSeaShip102 can be regarded as a complementary maritime long-term tracking benchmark rather than a comprehensive representation of all-weather maritime tracking conditions. Expanding the UAVSeaShip102 data set to cover more challenging illumination, weather, and target-interaction conditions will be an important direction in our future work.

4. Dynamic Multi-scale Fusion Transformer Tracker

Due to rapid target motion, drastic scale variations, and frequent occlusion in complex aerial scenarios, the UAV-based object tracking task usually encounters severe challenges. To address these challenges, we introduce a Dynamic Multi-scale Fusion Transformer Tracker, referred to as DMFTrack, for UAV-based single object tracking. In this section, we first provide a brief overview of the proposed DMFTrack. Then we present the Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM) in detail. Furthermore, the Adaptive Template Update Mechanism (ATUM) is described. Finally, we introduce the prediction head and loss function used for training.

4.1. Overview

Our proposed tracker adopts a compact Vision Transformer (ViT) backbone as the general feature extractor. The backbone is not treated as the primary methodological contribution. Instead, it provides a unified representation space for the proposed spatial and temporal adaptation mechanisms. As illustrated in Figure 4, the proposed DMFTrack

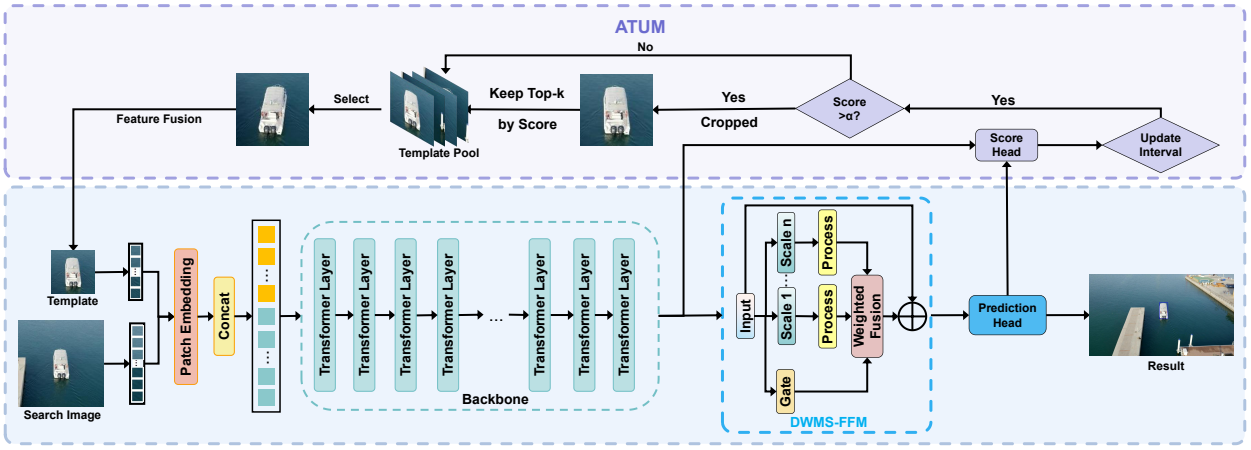


Figure 4: Overall architecture of the proposed DMFTrack, which comprises a single-stream backbone, a basic prediction head with DWMS-FFM (Novelty 1, Section 4.2) and a score head used for guiding the ATUM (Novelty 2, Section 4.3).

addresses two complementary sources of uncertainty in UAV-based long-term tracking. The Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM) is introduced between the backbone and prediction head to adaptively aggregate representations from multiple spatial scales. This design enables the tracker to dynamically balance fine-grained target details and broader contextual information according to the current input. On the other hand, the Adaptive Template Update Mechanism (ATUM) introduces a dedicated Score Head to estimate the reliability of the current template-search pair and selectively update the template memory. Consequently, the DWMS-FFM mainly addresses spatial-scale variation, whereas the ATUM focuses on temporal template degradation and long-term drift.

Given a template image $Z \in \mathbb{R}^{3\times H_z\times W_z}$ and a search image $S \in \mathbb{R}^{3\times H_s\times W_s}$, the backbone transformer encodes them as feature tokens. This process can be expressed as

$$X = \mathrm{E}(Z, S) = [X_z, X_s], \tag{1}$$

where $E(\cdot)$ denotes the patch embedding and transformer encoding process. The resulting spatial feature maps are further refined by the proposed DWMS-FFM, as foumulated by the equation below:

$$F = \text{DWMS-FFM}(X). \tag{2}$$

Finally, X and F are sent to a classic prediction head and a score head, which are used to obtain the tracking results and guide the adaptive template update, respectively.

4.2. Dynamically-Weighted Multi-scale Feature Fusion Module

Existing multi-scale fusion methods usually assign fixed coefficients to features across different receptive fields, which may limit adaptability in dynamically changing UAV-based tracking scenarios. However, the scale and contextual information of the targets will continuously vary due to the motion and viewpoint changes of the UAV in aerial tracking scenarios. Therefore, a fixed fusion strategy may fail to preserve the most informative representations for robust tracking. To enhance the spatial awareness and robustness of the search features, we propose a Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM), as shown in Figure 5. This module adaptively aggregates multi-scale features using a learnable gating mechanism. Specifically, given a batch of search feature maps $x \in \mathbb{R}^{B\times C\times H\times W}$, we first generate a set of scaled versions of x using bilinear interpolation. Each set of scaled feature maps is processed by a lightweight convolutional branch to extract scale-specific representations.

Let $\{s_1, s_2, ..., s_N\}$ denote the set of predefined scales. For each scale s_i, the corresponding feature map is computed as:

$$\begin{aligned} f_i &= \mathrm{Conv}_i(\mathrm{Resize}(x, s_i)), \\ &= \begin{cases} \mathrm{Conv}_i(x), & if\ s_i = 1.0, \\ \mathrm{Conv}_i(\mathrm{Interpolate}(x, H_i, W_i)), & otherwise, \end{cases} \end{aligned} \tag{3}$$

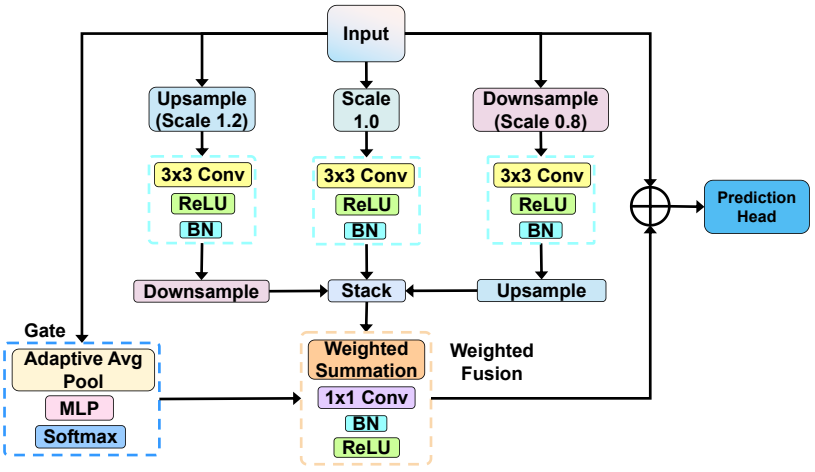


Figure 5: Architecture of the proposed Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM).

where each branch extracts complementary contextual cues. All resulting features are resized back to the same spatial resolution and stacked to form a multi-scale tensor:

$$F = [f_1, f_2, ..., f_N] \in \mathbb{R}^{B\times N\times C'\times H\times W}, \tag{4}$$

where $C' = C/r$ is the reduced channel dimension.

To dynamically fuse these multi-scale features, we design a dynamic gating network that predicts a set of attention weights based on the global context of the input features. The gating network consists of two fully-connected layers with the ReLU activation function, which is expressed as

$$w = \text{Softmax}(W_2 \cdot \text{ReLU}(W_1 \cdot (\text{GAP}(x) + b_1)) + b_2), \tag{5}$$

where $\text{GAP}(\cdot)$ denotes global average pooling, and $w_i \in \mathbb{R}^{B\times N}$ represents normalized attention weights across scales. The fused representation is obtained as:

$$f_{fused} = \sum_{i=1}^{N} w_i \cdot f_i, \tag{6}$$

Finally, the fused features are mapped back to the original channel dimension via a 1×1 convolution and added to the input features with a residual connection. This process can be formulated as

$$y = x + \text{Fusion}(f_{fused}). \tag{7}$$

The three scale factors, including 0.8, 1.0, and 1.2, are introduced to construct a compact coarse-to-fine spatial representation around the original feature resolution. The 1.0 branch preserves the original spatial representation and therefore serves as the reference branch without introducing additional interpolation. In contrast, the 0.8 branch first produces a coarser spatial representation. The reduced spatial sampling suppresses the sensitivity to fine local variations and allows the subsequent convolutional operations to aggregate information over a relatively broader spatial context. This branch is therefore beneficial when the target is strongly affected by background fluctuations or when contextual information is more important than precise local details. The 1.2 branch adopts a denser spatial sampling before the convolution operations. This branch does not create new image information through upsampling; instead, it provides a finer feature-processing grid, allowing the convolutional branch to preserve and process local structural cues at a finer spatial granularity. Such a representation is particularly useful for small or low-resolution targets and for accurate target localization. After scale-specific processing, all branches are resized back to the original resolution before the fusion operation is performed.

In particular, the values of 0.8 and 1.2 are selected as moderate and approximately symmetric perturbations around the original scale of 1.0. This design provides both a coarser contextual branch and a finer spatial-processing branch

Algorithm 1 Adaptive Template Update Mechanism

while 1<t<T **do**
  if $t \mod T_u = 0$ **and** $c_t > \tau$ **then**
    $\mathcal{T}_{\mathrm{t}} \leftarrow Preprocessor\big(Crop(I_t, \mathbf{b}_t)\big)$
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{(\mathcal{T}_{\mathrm{t}}, c_t)\}$
    sort $\mathcal{P}$ by c_i in descending order
    $\mathcal{P} \leftarrow \mathcal{P}[1..K_{\max}]$ {keep top-$K_{\max}$}
    $\mathcal{T}_{\mathrm{rand}} \leftarrow RandomSelect(\mathcal{P})$
    $\mathcal{T}_{online} \leftarrow \alpha \cdot \mathcal{T}_{online} + (1-\alpha) \cdot \mathcal{T}_{\mathrm{rand}}$
  end if
  return $\mathcal{P}$, $\mathcal{T}_{\mathrm{online}}$
end while

while keeping the scale range compact, thereby limiting interpolation distortion and additional computational cost. It should be noted that both 0.8 and 1.2 are not considered optimal dataset-specific values. Instead, they constitute a practical coarse-to-fine scale set that provides sufficient spatial diversity for dynamic fusion without introducing a large number of branches. The subsequent gating network determines the contribution of each scale according to the input feature content, rather than assigning a fixed preference to a predefined scale.

The key distinction of DWMS-FFM from conventional multi-scale fusion lies in the generation of scale weights. The three spatial representations are not combined with fixed coefficients. On the contrary, the gate network derives normalized weights directly from the current input features, making the contribution of each scale content-dependent. This design enables scale-adaptive feature emphasis, allowing the tracker to dynamically adjust its receptive field depending on the target size or motion patterns. As a result, DWMS-FFM enhances the ability of the ViT backbone to simultaneously capture fine-grained structures and large-context cues without increasing inference cost greatly.

4.3. Adaptive Template Update Mechanism

Traditional template update strategies often rely on fixed update intervals or heuristic exponential moving average mechanisms, which may gradually accumulate inaccurate target representations during long-term tracking. Once unreliable templates are introduced, tracking drift will further spread in the subsequent frames. To adapt to significant appearance variations of the target during long-term UAV-based tracking, we design an Adaptive Template Update Mechanism (ATUM), as illustrated in Algorithm 1. It selectively updates the template features based on the predicted confidence score and a learned update policy. Unlike traditional fixed-template strategies, which are vulnerable to drift under occlusion or deformation, our ATUM maintains a template pool and dynamically selects or updates templates during the tracking process.

We maintain a template pool $\mathcal{P} = \{(\mathcal{T}_1, c_1), (\mathcal{T}_2, c_2), ..., (\mathcal{T}_k, c_k)\}$ with a fixed capacity k, where $\mathcal{T}_i$ is a feature vector and $c_i \in [0, 1]$ is the confidence given by the Score Head. At the beginning of tracking, the initial template $(\mathcal{T}_0, c_0)$ is extracted from the first frame to guarantee a stable anchor. To explicitly estimate the reliability of the current target observation before updating the template memory, we employ a lightweight Score Head based on two successive cross-attention stages. Let $F_s \in \mathbb{R}^{C\times H\times W}$ denote the search feature maps and $F_t \in \mathbb{R}^{C\times H_t\times W_t}$ denote the template feature maps, where C is the dimension of feature channels, H and W are the spatial dimensions of the search feature maps, and H_t and W_t are the corresponding spatial dimensions of the template feature map. Given a predicted target bounding box b, we first extract the target-region representation from the search features using precise Region-of-Interest (RoI) pooling, which can be expressed as

$$R_s = \mathrm{PrRoIPool}(F_s, b), \tag{8}$$

where $\mathrm{PrRoIPool}(\cdot)$ denotes precise RoI pooling and $R_s \in \mathbb{R}^{C\times h_r\times w_r}$ represents the features corresponding to the currently predicted target region. The pooled features are then flattened into a sequence of search-region tokens, denoted by $X_s \in \mathbb{R}^{N_s\times C}$, where $N_s = h_r w_r$ is the number of pooled spatial tokens.

Similarly, the template features are flattened into a template-token sequence $X_t \in \mathbb{R}^{N_t\times C}$, where $N_t = H_t W_t$ denotes the number of template tokens. A learnable score token $q_0 \in \mathbb{R}^{1\times C}$ is introduced to summarize the reliability-related information. In the first cross-attention stage, the score token interacts with the current search-region representation, which can be formulated as

$$q_1 = \mathrm{LN}\left(q_0 + \mathrm{MHA}_s(q_0, X_s, X_s)\right), \tag{9}$$

where $\mathrm{MHA}_s(Q, K, V)$ denotes the multi-head attention mechanism with query Q, key K, and value V, and $\mathrm{LN}(\cdot)$ represents layer normalization. The result q_1 is a search-aware reliability representation obtained by aggregating evidence from the currently predicted target region.

In the second stage, the search-aware representation further interacts with the template representation. This process can be expressed as

$$q_2 = \mathrm{LN}\left(q_1 + \mathrm{MHA}_t(q_1, X_t, X_t)\right), \tag{10}$$

where X_t provides the stored target appearance information and q_2 is the resulting template-aware reliability representation. In essence, the first cross-attention stage focuses on the evidence contained in the current target observation, while the second stage incorporates the historical target appearance encoded by the template.

Finally, the refined score token is mapped to a scalar reliability score using the equation below

$$c_t = \mathrm{MLP}(q_2), \tag{11}$$

where $\mathrm{MLP}(\cdot)$ denotes the score-prediction Multilayer Perceptron (MLP) and c_t is the predicted reliability score of the current target observation.

At each template update interval T_u , we randomly select a historical template from the template pool and combine it with the current template using a weighted summation operation. This operation can be formulated as

$$\mathcal{T}_{online} = \alpha \cdot \mathcal{T}_{online} + (1 - \alpha) \cdot \mathrm{Random}(\mathcal{P}), \tag{12}$$

where $\mathcal{T}_{online}$ represents the template currently used for tracking, $\mathcal{P}$ denotes the pool of high-quality templates, and α stands for the weighting coefficients which control the balance between the high-quality templates and the current template. The weighted summation operation enhances the robustness of the tracker, which ensures that the tracker can both update information promptly and retain its original important features. Meanwhile, only storing high-quality templates and strictly limiting the size of the pool can also meet real-time requirements on the UAV. The contribution of the ATUM lies in the manner in which the estimated reliability is incorporated into the temporal template management process. Specifically, the reliability score is used to regulate the admission of the candidate template and guide the fusion of current reliable observations with existing template representations, rather than simply serving as a binary replacement criterion.

In our experiments, the template update interval T_u was set to 100, the confidence threshold τ was set to 0.9, the template pool size $K_{\max}$ was fixed at 3, and the update coefficient α was set to 0.9, unless otherwise specified. This confidence-guided update mechanism suppresses the accumulation of erroneous templates when encountering temporary appearance degradation, occlusion, and background interference, thereby improving long-term tracking stability.

4.4. Prediction Head and Training Loss

The prediction head is implemented as a corner detection head with DWMS-FFM. Following the corner detection head used in the previous work (Cui et al., 2022), we utilize a prediction head consisting of multiple Conv-BN-ReLU layers to directly estimate the bounding box of the target. The output tokens corresponding to the search image are first reshaped to a 2D spatial feature map and then fed into the prediction head. The head outputs a local offset $o \in [0, 1]^{2\times H_x/P\times W_x/P}$, a normalized bounding box size $s \in [0, 1]^{2\times H_x/P\times W_x/P}$, and a target classification score $p \in [0, 1]^{H_x/P\times W_x/P}$. The initial estimation of the target position depends on identifying the location with the highest classification score according to

$$(x_c, y_c) = \mathrm{argmax}_{(x,y)}\mathrm{P}(x, y). \tag{13}$$

The final target bounding box is estimated by$\{(x_t, y_t); (w, h)\} = \{(x_c, y_c) + o(x_c, y_c); s(x_c, y_c)\}$.

Following existing studies (Wu et al., 2025b; Cui et al., 2022), we conduct the offline training operation in two stages. Within the first stage, the modules used for the tracking task are trained. During the second stage, only the score head utilized for the confidence prediction task is trained. In particular, the score head is trained as a binary classifier to estimate the reliability of the template-search pair, in which the ground-truth labels are constructed based on the consistency between the template and search regions. Positive samples are generated by sampling templates and search frames from the same object instance, indicating reliable matching. In contrast, negative samples are constructed either

by sampling frames where the target is absent (e.g., occlusion or out-of-view) or by assigning mismatched bounding boxes.

For the tracking task, we adopt the weighted focal loss (Law and Deng, 2018) for classification, and a combination of L_1 loss and GIoU loss (Rezatofighi, Tsoi, Gwak, Sadeghian, Reid and Savarese, 2019) for bounding box regression. For the confidence predicting task, we use a binary cross-entropy with logits loss. The loss functions used for tracking prediction and confidence prediction are defined as

$$\mathcal{L}_{pred} = \mathcal{L}_{cls} + \lambda_{iou}\mathcal{L}_{iou} + \lambda_{L_1}\mathcal{L}_{L_1}, \tag{14}$$

$$\mathcal{L}_{conf} = \lambda_{score}\mathcal{L}_{score}, \tag{15}$$

respectively, where $\lambda_{iou} = 2$, $\lambda_{L_1} = 5$ and $\lambda_{score} = 1$ are the regularization parameters.

5. Experiments

In this section, we first introduce the settings used in our experiments. Then we report the experimental results obtained in this study. Furthermore, we evaluated different trackers by fine-tuning them using our UAVSeaShip102. Finally, the results derived in an ablation study are presented.

5.1. Experimental Settings

We will introduce the backbone network, data sets, performance metrics and implementation details that we used in our experiments.

5.1.1. Backbone

Following existing studies (Li et al., 2024b; Wu, Li, Liu, Wang, Yang, Ye, Zeng, Zhao and Li, 2025a; Xue et al., 2025), we utilized the pre-trained DeiT-Tiny (Touvron et al., 2021) model, which had achieved a good balance between tracking accuracy and computational efficiency, as the backbone network. We adopted a tracker on top of the proposed DMFTrack and this backbone, referred to as DMFTrack-DeiT. Specifically, the head of DMFTrack consists of a stack of four Conv-BN-ReLU layers. The search region and template sizes were set to 256×256 and 128×128 pixels, respectively.

5.1.2. Data Sets

To demonstrate the effectiveness of our method, both our UAVSeaShip102 data set and six publicly available UAV-acquired data sets, including DTB70 (Li and Yeung, 2017), UAV123, UAV123@10fps (Mueller et al., 2016), UAVTrack112, and UAVTrack112L (Fu et al., 2021a), were used for evaluation. Specifically, DTB70 consists of 70 UAV-acquired sequences, encompassing various challenges, such as UAV motion, cluttered scenes and objects with different sizes. UAV123 is a large-scale UAV-based tracking benchmark, which comprises 123 sequences with more than 112K frames. UAV123@10fps was constructed by down-sampling the original UAV123 benchmark from 30 FPS to 10 FPS to assess the impact of the camera capture speed on tracking performance. UAVTrack112 contains 112 challenging aerial scenarios, covering over 100K frames. As the largest existing long-term aerial tracking benchmark, UAVTrack112L includes over 60k frames.

5.1.3. Performance Metrics

We evaluated our DMFTracK and baseline trackers using four commonly-used performance metrics for UAV-based tracking, including Success Rate, Area Under the Curve (AUC), Precision (P), and Normalized Precision (P_{Norm}).

Following the One-Pass Evaluation (OPE) protocol, the success plot is generated based on the Intersection-over-Union (IoU) overlap between the predicted and the ground-truth bounding boxes. For the i-th frame, the IoU is computed as

$$IoU_i = \frac{|B_i^{pred} \cap B_i^{gt}|}{|B_i^{pred} \cup B_i^{gt}|}, \tag{16}$$

where B_i^{pred} and B_i^{gt} denote the predicted and ground-truth bounding boxes, respectively. Given an IoU threshold τ, the Success Rate is defined as

$$S(\tau) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(IoU_i \geq \tau), \tag{17}$$

where N denotes the total number of frames and $\mathbf{1}(\cdot)$ is the indicator function defined as

$$\mathbf{1}(x) = \begin{cases} 1, & x \text{ is true}, \\ 0, & x \text{ is false}. \end{cases} \tag{18}$$

The success plot is obtained by varying τ from 0 to 1. To provide a comprehensive evaluation across all overlap thresholds, the AUC score is computed as

$$AUC = \int_0^1 S(\tau)d\tau. \tag{19}$$

Unlike a single-threshold success score, AUC reflects the overall localization performance of a tracker over the entire range of IoU thresholds.

Precision evaluates the tracking accuracy from the perspective of target center localization. Specifically, the center location error for the i-th frame is defined as

$$d_i = \sqrt{(x_i^{pred} - x_i^{gt})^2 + (y_i^{pred} - y_i^{gt})^2}, \tag{20}$$

where (x_i^{pred}, y_i^{pred}) and (x_i^{gt}, y_i^{gt}) denote the centers of the predicted and ground-truth bounding boxes, respectively. Precision is then computed as

$$P = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(d_i \leq 20), \tag{21}$$

which measures the percentage of frames whose center location errors are within 20 pixels.

Since the same pixel error may have different impacts on targets of different sizes, Normalized Precision further normalizes the center location error using the width and height of the ground-truth bounding box. The normalized center location error is defined as

$$\hat{d}_i = \sqrt{(\frac{x_i^{pred} - x_i^{gt}}{w_i^{gt}})^2 + (\frac{y_i^{pred} - y_i^{gt}}{h_i^{gt}})^2}, \tag{22}$$

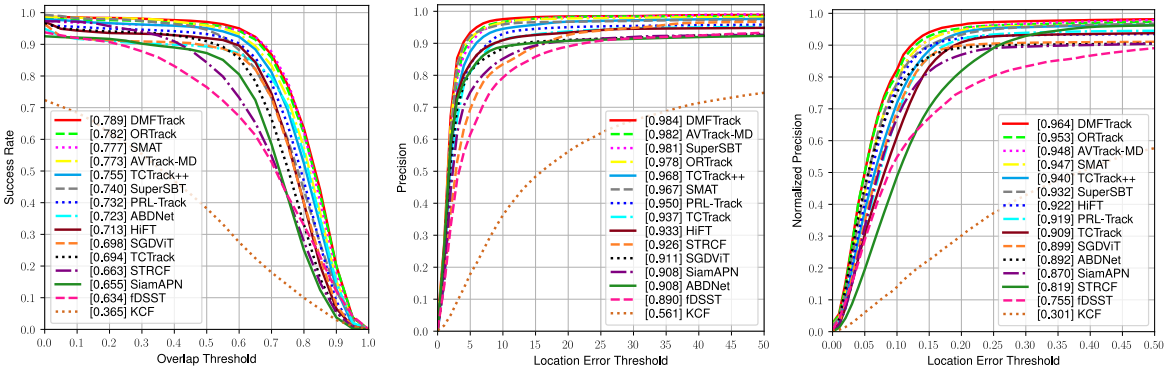
where w_i^{gt} and h_i^{gt} denote the width and height of the ground-truth bounding box, respectively. Following the TrackingNet (Müller, Bibi, Giancola, Alsubaihi and Ghanem, 2018) evaluation protocol, the normalized precision curve is first computed by varying the normalized location error threshold τ from 0 to 0.5 as

$$P_{Norm}(\tau) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{d}_i < \tau). \tag{23}$$

The final normalized precision score is obtained by averaging the normalized precision curve over all thresholds as

$$P_{Norm} = \frac{1}{M}\sum_{j=1}^{M}P_{Norm}(\tau_j), \tag{24}$$

where M denotes the number of sampled thresholds in the interval $[0, 0.5]$. Higher values of Success Rate, AUC, Precision (P), and Normalized Precision (P_{Norm}) indicate better tracking performance.



(a) Success (b) Precision (c) Normalized Precision

Figure 6: Comparison between the proposed DMFTrack and 14 baseline UAV-based trackers in three different performance metrics on our UAVSeaShip102.

5.1.4. Implementation Details

Following exiting protocols, we trained our models on four data sets, including LaSOT (Fan et al., 2019), COCO (Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár and Zitnick, 2014), TrackingNet (Müller et al., 2018), and GOT-10k (Huang et al., 2021). As performed for OSTrack (Ye et al., 2022), we used typical commonly-used data augmentation methods, such as horizontal flipping and brightness jittering. Following the existing study, we conducted the offline training operation in two stages, in which the network was trained for 300 epochs and 50 epochs, respectively. Both stages utilized 60k matching pairs per epoch.

Our tracker was implemented using Python 3.8 and PyTorch 1.7. With the AdamW optimizer (Loshchilov and Hutter, 2017), the learning rates were set to 4×10^{-4} for the prediction head. The learning rates dropped by a factor of 10 after 240 epochs. Within the first stage, the entire network except for the score head was trained while the score head was kept frozen. During the second stage, only the score head was trained and the other modules of the network were kept frozen. The model training was conducted on an NVIDIA L40 Graphics Processing Unit (GPU), and the batch size was fixed at 32.

We followed the common practice and employed the Hamming window to incorporate positional priors. Specifically, a penalty was applied to the classification map p using the Hanning window. Then the box corresponding to the highest score was selected as the prediction bounding box.

5.2. Experimental Results

5.2.1. Quantitative Evaluation

We first tested the DMFTrack model that we trained using the four training data sets on our UAVSeaShip102 data set. That is, UAVSeaShip102 was not used for training but was utilized only for evaluation, ensuring a fair assessment of generalization. Although the maritime data was not involved in the training process, maritime UAV-based tracking shares several common challenges with generic UAV-based tracking, such as long-term appearance variations, large scale changes and target deformation. Among these challenges, large scale variations can be effectively handled by the proposed DWMS-FFM, while long-term appearance changes can be addressed by the ATUM, which continuously maintains discriminative template representations during the tracking process. The results obtained using our method and 14 baseline approaches are shown in Figure 6. As can be seen, our DMFTrack performs the best in terms of the three metrics. The results suggest that the proposed spatial and temporal adaptation mechanisms can effectively transfer the pre-trained model to unseen UAV-based maritime sequences without fine-tuning in the maritime domain.

Then we compared DMFTrack against 22 state-of-the-art lightweight trackers. Unlike generic computer vision tasks such as object detection and image segmentation, UAV-based single object tracking involves continuous target localization and matching under long-term appearance variations, scale instability, and background interference. Therefore, evaluation of UAV-based tracking methods on specialized UAV tracking benchmarks is essential to accurately assess tracking robustness and real-time applicability. Table 4 shows the precision and AUC values produced by our method and the baselines on the DTB70 (Li and Yeung, 2017), UAV123 (Mueller et al., 2016), UAV123@10fps

Category	Method	Venue	DTB70		UAV123		UAV123@10fps		UAVTrack112		UAVTrack112L		Avg.	
			Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC
DCF-Based	fDSST (Danelljan, Häger, Khan and Felsberg, 2017b)	TPAMI 17	53.4	35.7	58.3	40.5	51.6	37.9	56.8	39.1	49.1	30.6	53.8	36.8
	ECO_HC (Danelljan et al., 2017a)	CVPR 17	63.5	44.8	71.0	49.6	64.0	46.8	68.6	47.4	64.8	41.7	66.4	46.1
	MCCT_H (Wang, Zhou, Tian, Hong, Wang and Li, 2018)	CVPR 18	60.4	40.5	65.9	45.7	59.6	43.4	63.4	43.6	58.5	37.5	61.6	42.1
	STRCF (Lan, Yang, Zhang and Yuen, 2021)	CVPR 18	64.9	43.7	68.1	48.1	62.7	45.7	63.5	43.0	60.9	36.0	64.0	43.3
	ARCF (Huang, Fu, Li, Lin and Lu, 2019)	ICCV 19	69.4	47.2	67.1	46.8	66.6	47.3	67.2	45.6	64.0	39.3	66.9	45.2
	AutoTrack (Li, Fu, Ding, Huang and Lu, 2020)	CVPR 20	71.6	47.8	68.9	47.2	67.1	47.7	69.4	46.5	67.5	40.2	68.9	45.9
CNN-Based	HiFT (Cao, Fu, Ye, Li and Li, 2021)	ICCV 21	80.2	59.4	78.7	59.0	74.9	57.0	74.2	57.0	73.4	55.1	76.3	57.5
	SiamAPN (Fu, Cao, Li, Ye and Feng, 2021b)	ICRA 21	78.4	58.5	76.5	57.5	75.2	56.6	80.9	61.6	78.0	57.7	77.8	58.4
	TCTrack (Cao, Huang, Pan, Zhang, Liu and Fu, 2022)	CVPR 22	81.1	61.9	77.3	60.4	75.1	58.8	76.6	59.4	78.6	58.3	77.7	59.8
	TCTrack++ (Cao, Huang, Pan, Zhang, Liu and Fu, 2023)	TPAMI 23	80.4	61.7	74.4	58.8	78.2	60.1	81.5	64.0	82.0	62.4	79.3	61.4
	SGDViT (Yao, Fu, Li, Zheng and Ye, 2023)	ICRA 23	78.5	60.4	75.4	57.5	**86.3**	*66.1*	77.0	60.0	73.3	54.7	78.1	59.7
	ABDNet (Zuo, Fu, Li, Lu, Li and Feng, 2023)	RAL 23	76.8	59.6	79.3	60.7	77.3	59.1	77.8	61.0	78.1	59.2	77.9	59.9
	PRL-Track (Fu, Lei, Zuo, Yao, Zheng and Pan, 2024)	IROS 24	79.5	60.6	79.1	59.3	74.1	58.6	81.7	62.7	80.6	59.9	79.0	60.2
ViT-Based	LiteTrack–B4 (Yan, Peng, Wu, Wang, Fu and Lu, 2021b)	ICRA 24	82.5	63.9	84.2	65.9	83.1	64.0	*82.9*	66.9	78.2	61.4	82.2	64.4
	SMAT (Gopal and Amer, 2024)	WACV 24	81.9	63.8	81.8	64.6	80.4	63.5	82.2	65.3	76.5	60.8	80.6	63.6
	LightFC-vit (Li, Wang, Wu, Liu and Li, 2024c)	KBS 24	82.8	64.0	84.2	65.5	81.3	63.7	81.3	65.7	78.7	62.2	81.7	64.2
	SuperSBT-Light (Xie, Yang, Wang, Chu, Cao, Ma and Zeng, 2024)	TPAMI 24	*84.5*	65.4	*85.0*	*67.2*	82.1	65.3	82.9	66.1	79.9	62.1	82.9	65.2
	AVTrack-DeiT (Li et al., 2024b)	ICML24	84.3	65.0	84.8	66.8	83.2	65.8	80.3	65.4	78.2	62.7	82.2	65.1
	AVTrack-MD-DeiT (Wu et al., 2025a)	TCSVT 25	84.0	65.2	82.6	65.2	83.3	65.5	81.8	66.4	78.8	62.7	82.1	65.0
	ORTrack-DeiT (Wu et al., 2025b)	CVPR 25	**86.2**	**66.4**	84.3	66.4	82.7	65.2	81.3	65.6	78.6	62.7	82.6	65.3
	ORTrack-D-DeiT (Wu et al., 2025b)	CVPR 25	83.7	65.1	84.0	66.1	82.3	65.1	81.8	66.0	79.8	63.0	82.3	65.1
	SGLATrack-DeiT (Xue et al., 2025)	CVPR 25	84.4	65.1	84.9	66.9	82.6	65.5	82.8	**67.5**	79.2	*64.0*	82.8	*65.8*
	DMFTracK-DeiT	Ours	84.2	*65.5*	**86.1**	**67.8**	*84.7*	**67.0**	**83.1**	*67.3*	**82.7**	**66.1**	**84.1**	**66.7**

Table 4
Comparison between our method and 22 state-of-the-art lightweight trackers in terms of the precision and AUC metrics on the test sets of DTB70 (Li and Yeung, 2017), UAV123 (Mueller et al., 2016), UAV123@10fps (Mueller et al., 2016), UAVTrack112 (Fu et al., 2021a) and UAVTrack112L (Fu et al., 2021a). In addition, the average metric values computed across the five data sets are presented in the column of "Avg.". Here, the top three results derived in each case are highlighted in the **red**, *blue* and cyan fonts, respectively.

(Mueller et al., 2016), UAVTrack112 (Fu et al., 2021a) and UAVTrack112L (Fu et al., 2021a) data sets. It can be observed that our DMFTrack achieves the best AUC scores on the UAV123 (Mueller et al., 2016), UAVTrack112 (Fu et al., 2021a), and UAVTrack112L (Fu et al., 2021a) data sets while normally ranking the top-three in both metrics across all the data sets.

Although DMFTrack does not achieve the best result on the DTB70 data set, it still maintains competitive tracking accuracy compared to recent Transformer-based trackers. This finding should be attributed to the characteristics of the DTB70 data set. Compared to the UAVTrack112L and UAVSeaShip102 data sets, most DTB70 sequences are relatively short and contain fewer severe long-term appearance variations. Consequently, the proposed ATUM, which is specifically designed to alleviate template degradation during long-term tracking, cannot fully manifest its advantages. In addition, extreme target scale variations occur less frequently in the DTB70 data set, reducing the contribution of the proposed DWMS-FFM. Nevertheless, DMFTrack still achieves competitive performance on this data set while demonstrating more significant improvements on the data sets containing longer sequences and larger appearance variations, validating the effectiveness of the proposed modules in more challenging UAV-based tracking scenarios.

It should be noted that the above results show that the pre-trained DMFTrack can be directly transferred to the unseen maritime domain and achieves competitive performance without fine-tuning in the maritime domain. This observation supports the applicability of the proposed spatial-scale and temporal-template adaptation mechanisms across different UAV-based tracking domains.

5.2.2. Qualitative Evaluation

In Figure 7, we further visualized the results derived using our method and six baselines on six challenging sequences included in our data set. It can be seen that our DMFTrack performs well in different scenarios. Compared with the baselines, DMFTrack generally produces bounding boxes that are more closely aligned with the ground-truth target and achieves more stable localizations across consecutive frames. When the size or position of the target undergoes significant changes or when there are similar targets nearby, in particular, the boxes predicted using the baselines show relatively larger deviations. In contrast, DMFTrack can better preserve the target location and extent. These observations demonstrate that DMFTrack can produce consistent target localizations under varying visual conditions, further supporting its robustness in UAV-based maritime object tracking.

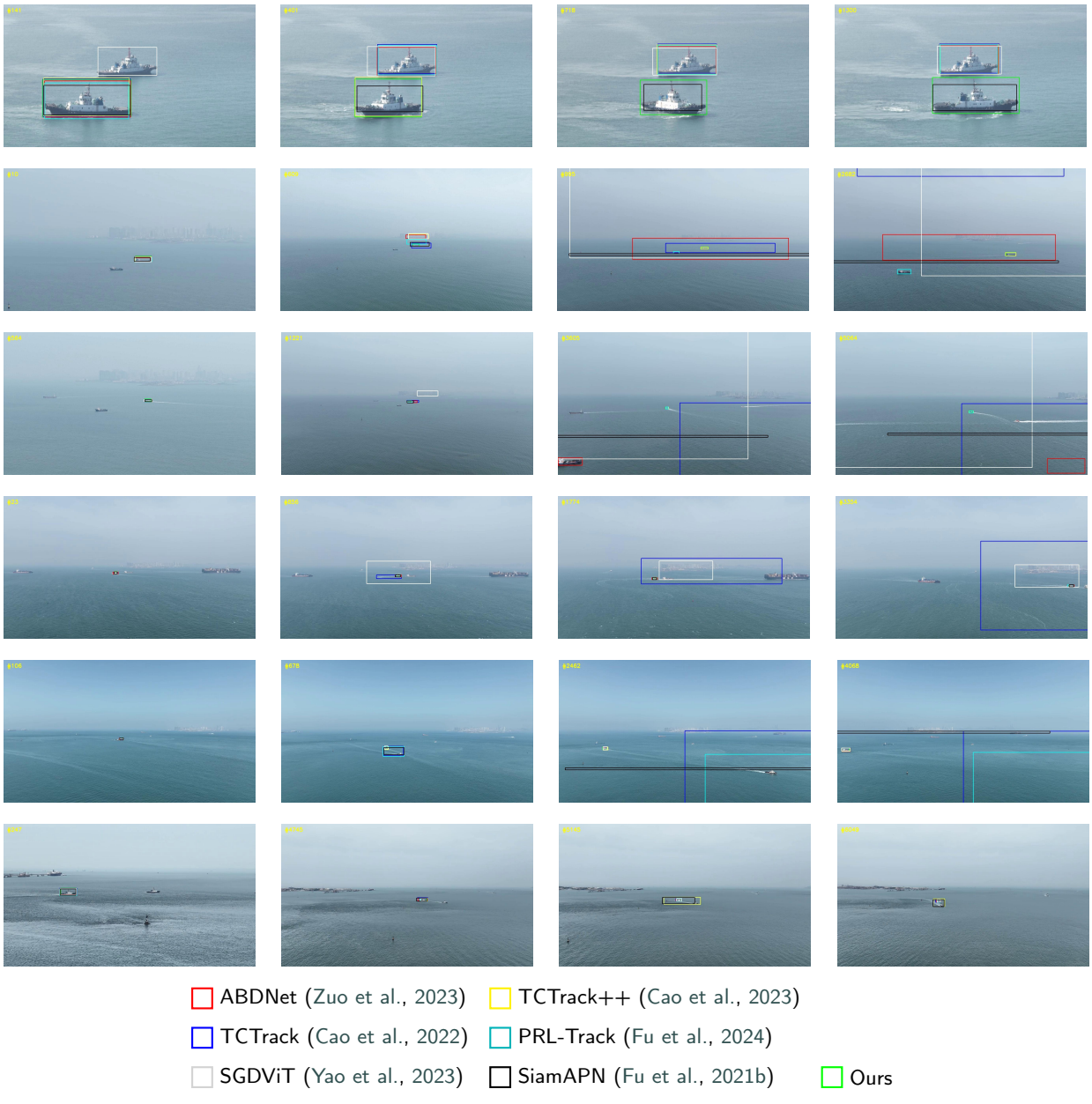


Figure 7: Visualization of the results of our tracker and six state-of-the-art baselines on six challenging sequences contained in our UAVSeaShip102 data set, including Ship0010, Ship0029, Ship0030, Ship0034, Ship0035 and Ship0043 (from top to bottom).

5.2.3. Evaluation in Challenging Scenarios

To better analyze the effectiveness of DMFTrack under extreme UAV-based tracking conditions, we further evaluated it together with six baseline methods in five challenging scenarios, such as fast motion, low resolution, and long-term tracking, contained in UAVTrack112 (Fu et al., 2021a). As reported in Table 5, DMFTrack achieves the largest improvement in the Scale Variation scenario. This improvement should mainly benefit from the proposed DWMS-FFM, which aggregates multi-scale contextual information through dynamically weighted feature fusion. Consequently, the tracker can better adapt to the significant changes in target size commonly encountered in UAV-based tracking. For the Long-Term Tracking scenario, DMFTrack also achieves competitive performance. This can

Method	Fast Motion		Low Resolution		Long-Term Tracking		Aspect Ratio Change		Scale Variation	
	Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC
ABDNet(Zuo et al., 2023)	64.3	59.2	70.1	62.4	76.4	71.0	73.8	69.4	75.3	69.6
SGDViT(Yao et al., 2023)	58.1	53.6	65.3	57.1	71.6	66.1	68.5	64.3	70.3	64.5
SiamAPN(Fu et al., 2021b)	62.4	55.9	71.5	60.6	76.2	69.6	73.8	67.3	75.4	68.2
LightFC-vit(Li et al., 2024c)	65.8	61.9	72.8	66.2	78.7	74.1	75.8	72.3	77.7	72.9
ORTrack(Wu et al., 2025b)	*66.7*	*64.6*	72.0	*66.3*	78.6	*74.7*	75.6	*72.6*	77.6	*73.5*
PRL-Track(Fu et al., 2024)	66.6	63.0	*73.9*	64.7	*78.9*	72.9	*76.0*	71.9	*77.9*	71.6
DMFTrack-DeiT (Ours)	**70.8**	**68.7**	**77.7**	**73.2**	**82.6**	**79.6**	**80.8**	**79.2**	**81.8**	**78.6**

Table 5
Comparison between the proposed DMFTracK and six baseline methods in five challenging scenarios contained in UAVTrack112L (Fu et al., 2021a). Here, the top two results are highlighted in the **red** and *blue* fonts, respectively.

be attributed to the proposed ATUM, which dynamically updates template representations according to tracking confidence, effectively reducing template contamination and improving robustness against appearance changes. Within the Fast Motion scenario, competitive performance demonstrates that the combination of dynamic template updating and multi-scale feature fusion enables the tracker to maintain reliable target localization even under rapid camera motion.

5.2.4. Performance Analysis

In addition, we assessed the performance of our DMFTrack in terms of computational complexity (Floating-Point Operations or FLOPs), number of parameters and inference speed. As reported in Table 6, DMFTrack required 2.5 GMac computational complexity and achieved 117.1 FPS using our experimental setup. The efficiency results of the compared trackers were collected from their corresponding original publications, whereas the DMFTrack results were measured using our setup. Since these values may have been obtained under different hardware, software environments, input configurations, and profiling protocols, they were provided primarily as reference indicators rather than strictly controlled measurements for direct ranking.

To analyze the efficiency of the proposed components, we separately profile the computational overhead of the DWMS-FFM and ATUM. For fair comparison, the baseline, individual modules, and full DMFTrack were evaluated under the same input configuration and hardware/software platforms. As reported in Table 7, the DWMS-FFM introduces a limited increase in number of parameters and computational complexity due to its lightweight three-scale convolutional branches and compact gating network. The ATUM introduces additional overhead mainly from the Score Head and the two-stage template–search interaction. Nevertheless, the results reported indicate that our DMFTrack maintains a relatively low computational cost while achieving a real-time inference speed. It is suggested that our tracker achieves a favorable balance between tracking performance and computational cost.

To further evaluate the practical deployment capability of the DMFTrack, we conducted an additional inference test on an NVIDIA Jetson Nano edge platform. As shown in Table 8, our DMFTrack achieved an inference speed of 27.82 FPS during the test process. We also monitored a system memory consumption in the 15W power mode. To be specific, the baseline RAM usage was 2,254 MB, while the average and peak RAM usage reached 3,064.65 MB and 3,293 MB, respectively. The peak increase over the baseline was 1,039 MB. These measurements provide an additional assessment of the memory requirements associated with running our DMFTrack on an embedded platform. It should be noted that the memory values reported indicate the RAM usage of the entired system rather than the memory exclusively allocated to the DMFTrack. The results provide an empirical evidence of the feasibility of running the DMFTrack on resource-constrained embedded hardware, while also highlighting the importance of considering system-level memory consumption in practical deployment. These results demonstrate that our DMFTrack can run on resource-constrained embedded hardware and maintain a practical inference speed.

5.2.5. Summary

Extensive experiments on our UAVSeaShip102 data set and three publicly available UAV-acquired benchmarks, including UAV123 (Mueller et al., 2016), DTB70 (Li and Yeung, 2017) and UAVTrack112 (Fu et al., 2021a), demonstrate that our approach normally achieved superior performance in terms of accuracy, robustness, and efficiency, compared to not only the classical DCF-based but also the state-of-the-art CNN-based and Transformer-based trackers. The results derived across UAVSeaShip102 and existing UAV-based tracking benchmarks indicate that the proposed

Method	Venue	FLOPs (GMac)	Params (M)	Inference Speed (FPS)
HiFT (Cao et al., 2021)	ICCV 21	7.2	9.9	*160.3*
SiamAPN (Fu et al., 2021b)	ICRA 21	7.9	14.5	**194.4**
TCTrack++ (Cao et al., 2023)	TPAMI 23	8.8	17.6	125.6
SGDViT (Yao et al., 2023)	ICRA 23	11.3	23.3	110.5
ABDNet (Zuo et al., 2023)	RAL 23	8.3	12.3	130.2
PRL-Track (Fu et al., 2024)	IROS 24	7.4	12.1	135.6
LiteTrack (Yan et al., 2021b)	ICRA 24	6.8	26.2	119.7
SMAT (Gopal and Amer, 2024)	WACV 24	*3.2*	**8.6**	124.2
SuperSBT (Xie et al., 2024)	TPAMI 24	10.4	21.5	121.7
DMFTracK-DeiT	Ours	**2.5**	**8.6**	117.1

Table 6
Comparison between the proposed DMFTracK and nine baseline methods in terms of computational complexity (i.e., FLOPs) , the number of parameters and inference speed. Given a metric, the top three values are highlighted in the **red**, *blue* and cyan fonts, respectively.

Model	Params (M)	FLOPs (GMac)	FPS
Baseline	7.98	2.39	108.9
Baseline + DWMS-FFM	8.25	2.45	73.4
Baseline + ATUM	8.35	2.40	81.5
DMFTrack	8.6	2.46	117.1

Table 7
Comparison between our DMFTrack and the baseline and its two variants in computational overhead.

Metric	Value
Embedded platform	NVIDIA Jetson Nano
Power mode	15 W
FPS	27.82
Baseline system RAM usage	2,254 MB
Average system RAM usage	3,064.65 MB
Peak system RAM usage	3,293 MB
Peak RAM increase over baseline	1,039 MB

Table 8
Performance analysis of our DMFTrack conducted on an NVIDIA Jetson Nano toolkit.

framework can be effectively transferred across different UAV-based tracking domains. In particular, the performance achieved in the unseen UAVSeaShip102 domain demonstrates the cross-domain transferability of the proposed spatial-scale and temporal-template adaptation mechanisms. These findings indicate the broader applicability of the framework.

5.3. Evaluation by Fine-Tuning a Tracker Using UAVSeaShip102

To further investigate the role of UAVSeaShip102 in adapting trackers to maritime scenarios, we conducted an additional fine-tuning experiment using this data set. Unlike the cross-domain evaluation reported above, in which the trackers trained on general tracking data sets were directly evaluated on UAVSeaShip102, this experiment allowed the trackers to learn from a designated subset of UAVSeaShip102. The 102 sequences were divided at the sequence level into a training set of 25 sequences, a validation set of 5 sequences, and a test set of 72 sequences. Since this protocol separates the sequences utilized for model adaptation from those used for testing, it enables a more controlled evaluation of performance variations after the model has been fine-tuned on maritime data. The results were assessed using the same evaluation metrics as those utilized in the preceding experiments.

In Table 9, we compare the performance of different trackers on the 72 test sequences derived without and with fine-tuning. The results obtained without fine-tuning characterize the transferability of the pre-trained trackers to the maritime domain, whereas the results derived using fine-tuning reflect the performance of these trackers after supervised adaptation has been performed using the maritime training sequences. The differences between the two

Tracker	$AUC^{w/o}$	$AUC^{w/}$	ΔAUC	$P^{w/o}$	$P^{w/}$	ΔP	$P^{w/o}_{Norm}$	$P^{w/}_{Norm}$	ΔP_{Norm}
TCTrack (Cao et al., 2022)	68.9	69.6	+0.7	95.3	88.7	-6.6	91.4	87.2	-4.2
TCTrack++ (Cao et al., 2023)	75.1	78.1	+3.0	96.6	96.8	+0.2	93.5	95.4	**+1.9**
ABDNet (Zuo et al., 2023)	72.4	74.3	+1.9	91.8	91.0	-0.8	89.9	90.0	+0.1
PRL-Track (Fu et al., 2024)	72.6	72.7	+0.1	95.4	90.9	-4.5	92.3	89.2	-3.1
SMAT (Gopal and Amer, 2024)	77.0	78.6	+1.6	95.9	98.0	**+2.1**	94.1	95.4	+1.3
ORTrack (Wu et al., 2025b)	77.2	78.9	+1.7	97.2	98.5	+1.3	94.8	95.5	+0.7
AVTrack-MD (Wu et al., 2025a)	76.4	78.9	+2.5	**98.4**	98.5	+0.1	94.5	95.7	+1.2
DMFTrack	**78.4**	**83.7**	**+5.3**	98.3	**98.6**	+0.3	**96.6**	**98.0**	+1.4

Table 9
Comparison between the performance of different trackers achieved without and with fine-tuning on UAVSeaShip102. The best result is indicated in the red, bold fonts in terms of each metric.

DWMS-FFM	ATUM	AUC	P	P_{Norm}
✗	✗	77.1	97.4	94.6
✓	✗	78.1	97.5	94.8
✗	✓	77.0	97.9	96.0
✓	✓	**78.9**	**98.4**	**96.4**

Table 10
The effect of the DWMS-FFM and ATUM on the proposed DMFTrack when our UAVSeaShip102 data set was used.

sets of results quantify the performance gain induced by fine-tuning. It can be seen that our DMFTrack achieved an AUC value of 78.4 without fine-tuning and a value of 83.7 with fine-tuning. Meanwhile, the Precision and Normalized Precision values obtained increased from 98.3 and 96.6 to 98.6 and 98.0, respectively. In addition, seven baseline trackers usually produced performance gains using fine-tuning. These findings suggest that fine-tuning is beneficial for adapting a pre-trained tracker to the maritime domain.

5.4. Ablation Study

To validate the effectiveness of each component in the proposed DMFTrack, we conducted a series of ablation experiments on our UAVSeaShip102 data set. Table 10 reports the results obtained using our DMFTrack and its three variants derived by removing the DWMS-FFM and/or the ATUM.

5.4.1. Effect of the DWMS-FFM

To investigate the impact of the proposed DWMS-FFM, we compared the results obtained with and without the DWMS-FFM. When both the DWMS-FFM and the ATUM were removed, the AUC, precision and normalized precision values derived were 77.1%, 97.4% and 94.6%, respectively, as shown in Table 10. In contrast, the three values increased to 78.1%, 97.5% and 94.8%, respectively, when only the ATUM was removed. The performance gains demonstrate that the dynamic gating-based multi-scale fusion strategy is useful for achieving the better feature representation. Specifically, the DWMS-FFM enables the tracker to adaptively emphasize scale-relevant features, enabling more robust performance during strong perspective variations or when the size of target vessels changes rapidly in the UAV-acquired sequences.

These results suggest that our DWMS-FFM is able to capture richer semantic–spatial representations by fusing multi-scale features in a dynamic and context-aware manner, which is particularly beneficial in the UAV-based tracking task, where targets often experience large-scale fluctuations.

5.4.2. Effect of the ATUM

We then examined the effect of the ATUM, which uses the Score Head to evaluate the reliability of template updates. As reported in Table 10, the AUC, precision and normalized precision values achieved by our model were

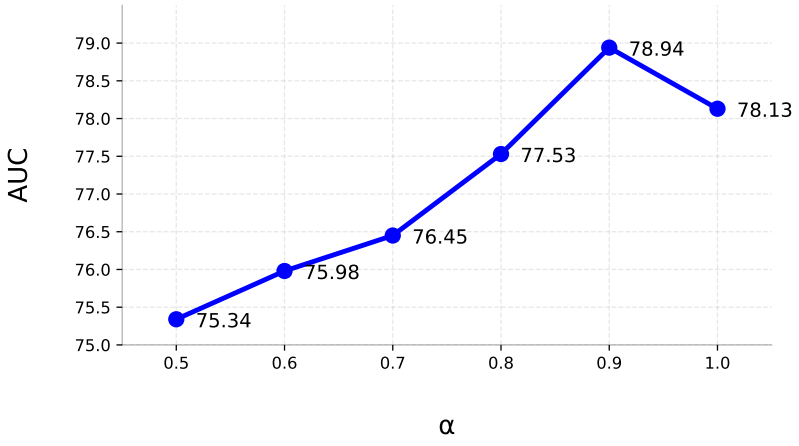


Figure 8: The AUC values produced by our method when different α values were used for the ATUM.

77.0%, 97.9% and 96.0%, respectively, when the ATUM was retained while the DWMS-FFM was removed. Compared with the three digits: 77.1%, 97.4% and 94.6% produced by our method when both the DWMS-FFM and the ATUM were removed, the utilization of the ATUM achieved moderate improvements, indicating that reliable template updates mitigate template degradation and enhance long-term stability.

5.4.3. Effect of the α Value

The gains of the ATUM will become more pronounced when sequences, which suffer from partial occlusion, illumination variation, or temporarily blurred frames, are processed. In these cases, naive template update strategies often lead to drift. Therefore, the choice of the value of α is key to the robustness of tracking. To investigate the impact of this parameter on our DMFTrack, different values of α were examined on the established UAV-based tracking benchmarks while the other settings were kept unchanged. The results are shown in Figure 8. It can be seen that the tracker tends to over-rely on historical templates when the value of α is excessively small, whereas larger α values reduce the adaptability of online template updating. In particular, the highest AUC value 78.94% was produced when the value of α was set to 0.9, which balanced the use of the current and historical templates. In addition, the similar trend was observed on the DTB70 (Li and Yeung, 2017) data set. As a result, $\alpha = 0.9$ was selected as the default setting.

5.4.4. The Joint Effect of the DWMS-FFM and ATUM

The results reported in Table 10 demonstrate that the DWMS-FFM and ATUM provide complementary improvements rather than redundant gains. Removal of the DWMS-FFM reduces the AUC from 78.9% to 77.0%, indicating the importance of adaptive multi-scale representation for spatially varying targets. In contrast, removing the ATUM mainly affects precision-oriented measures, with precision and normalized precision values decreasing from 98.4% and 96.4% to 97.5% and 94.8%, respectively. This difference is consistent with their intended roles. That is, the DWMS-FFM primarily improves spatial representation under scale variation, whereas the ATUM stabilizes the target representation over time and reduces the influence of unreliable observations. When both mechanisms are enabled, the tracker achieves the best overall performance, suggesting that spatial adaptation and temporal adaptation are complementary rather than redundant. The visualization results produced in the two cases are shown in Figure 9. As can be observed, our tracker obtained both strong discriminative capability and long-term robustness when the DWMS-FFM and ATUM were used together.

5.4.5. The Sensitivity of DMFTrack to the Backbone

To assess the sensitivity of the proposed DMFTrack to the backbone network, we further evaluated it using two additional Transformer-based backbones, including ViT-Tiny (Dosovitskiy et al., 2021) and EVA-Tiny (Fang, Wang, Xie, Sun, Wu, Wang, Huang, Wang and Cao, 2023). As shown in Table 11, the proposed framework remains effective when different compact Transformer backbones are employed. Although DeiT-Tiny (Touvron et al., 2021) achieved the

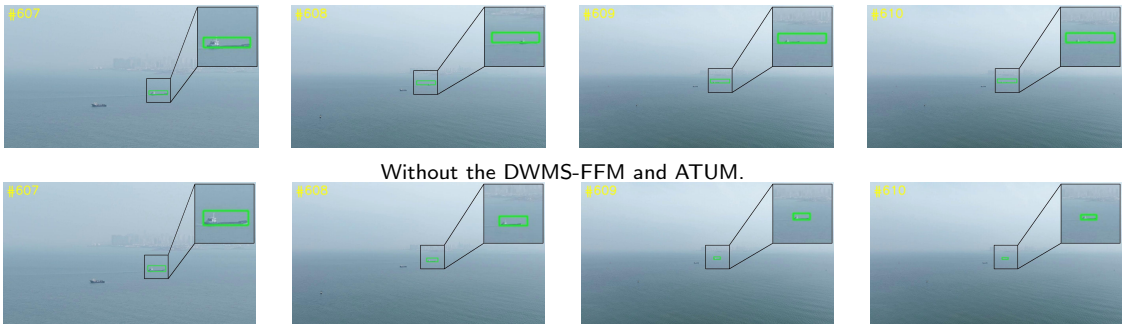


Figure 9: Comparison between the results obtained using our DMFTrack using both the DWMS-FFM and ATUM and without using these modules on the sequence "ship0029" in the UAVSeaShip102 data set.

Backbone	DTB70		UAV123		UAV123@10fps		UAVTrack112		UAVTrack112L		Avg.	
	Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC
DeiT-Tiny (Default)	**84.2**	**65.5**	**86.1**	**67.8**	84.7	**67.0**	**83.1**	**67.3**	**82.7**	**66.1**	**84.1**	**66.7**
ViT-Tiny	82.9	63.4	82.0	65.7	**85.3**	66.7	82.6	66.4	80.9	64.8	82.7	65.4
EVA-Tiny	81.4	63.6	84.3	66.0	82.7	64.8	81.2	65.2	77.6	61.5	81.5	64.2

Table 11
Comparison between the default DeiT-Tiny (Touvron et al., 2021) backbone and two Transform-based backbones, including ViT-Tiny (Dosovitskiy et al., 2021) and EVA-Tiny (Fang et al., 2023), together with the proposed DMFTrack on the test sets of DTB70 (Li and Yeung, 2017), UAV123 (Mueller et al., 2016), UAV123@10fps (Mueller et al., 2016), UAVTrack112 (Fu et al., 2021a) and UAVTrack112L (Fu et al., 2021a). In addition, the average metric values computed across the five data sets are presented. The best result in each column is highlighted in the **red** fonts.

best average performance among the three configurations, the performance remained relatively stable when replaced by ViT-Tiny (Dosovitskiy et al., 2021) or EVA-Tiny (Fang et al., 2023). This observation indicates that the effectiveness of DMFTrack does not originate solely from the choice of a particular off-the-shelf Transformer backbone. Instead, the proposed DWMS-FFM and ATUM provide task-oriented adaptation mechanisms that can be incorporated into different Transformer-based backbones.

6. Conclusion

In this paper, we first collected a UAV-acquired single object tracking data set, i.e., UAVSeaShip102, which contains 102 long-term maritime tracking sequences captured under challenging aerial and maritime conditions. Although this data set still has certain limitations, it offers a valuable benchmark for evaluating the robustness of trackers in real-world sea-surface scenarios and complements existing UAV-based tracking data sets.

To address the challenges that UAV-based long-term object tracking encounters, we further propose a lightweight Transformer-based tracking framework, termed Dynamic Multi-scale Fusion Transformer Tracker (DMFTrack). This framework comprises two complementary components, i.e., a Dynamically-Weighted Multi-scale Feature Fusion Module (DWMS-FFM) and an Adaptive Template Update Mechanism (ATUM), rather than relying on static feature aggregation and heuristic template updating strategies. The DWMS-FFM enables adaptive multi-scale feature aggregation through a dynamic gating strategy, allowing the tracker to flexibly emphasize scale-specific representations according to scene context. This design improves the ability of the model to make a trade-off between fine-grained localization accuracy and robust contextual understanding. On the other hand, the ATUM, driven by a confidence-aware Score Head, provides an effective strategy for stable template adaptation. Since the template will only be updated when its reliability is high, our method can mitigate model drift and maintain consistent tracking performance over long sequences.

Extensive experiments were performed on four UAV-based tracking benchmarks, including UAVSeaShip102, UAV123, DTB70, and UAVTrack112. The experimental results demonstrated that our DMFTrack achieved superior performance, compared to either classical trackers or state-of-the-art CNN-based and Transformer-based trackers, particularly in scenes with significant scale variations and long-term tracking. The ablation experiments further validated the effectiveness of dynamic multi-scale fusion and confidence-guided template updating strategies on improving both the accuracy and robustness of our model.

CRediT authorship contribution statement

Xinyue Wang: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. **Jingjing Xiao:** Formal analysis, Methodology, Writing - Review & Editing. **Xinghui Dong:** Conceptualization, Funding acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing - Review & Editing.

References

Bai, X., Zhang, Z., Xu, X., 2025. Advanced multispectral detection of maritime targets for unmanned ocean surveillance. Ocean Engineering 329, 121185. URL: https://www.sciencedirect.com/science/article/pii/S0029801825008984, doi:https://doi.org/10.1016/j.oceaneng.2025.121185.

Bhat, G., Danelljan, M., Gool, L.V., Timofte, R., 2019. Learning discriminative model prediction for tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Cao, Z., Fu, C., Ye, J., Li, B., Li, Y., 2021. Hift: Hierarchical feature transformer for aerial tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15457–15466.

Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., Fu, C., 2022. Tctrack: Temporal contexts for aerial tracking, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14778–14788. doi:10.1109/CVPR52688.2022.01438.

Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., Fu, C., 2023. Towards real-world visual tracking with temporal contexts. IEEE Trans. Pattern Anal. Mach. Intell. 45, 15834–15849. URL: https://doi.org/10.1109/TPAMI.2023.3307174, doi:10.1109/TPAMI.2023.3307174.

Cen, M., Jung, C., 2018. Fully convolutional siamese fusion networks for object tracking, in: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3718–3722. doi:10.1109/ICIP.2018.8451102.

Chen, C., Liang, X., Zhang, Z., Liu, D., Yu, C., Li, W., 2025. Cooperative game method of heterogeneous unmanned surface vehicles based on distributed decision-making framework. Ocean Engineering 338, 122048.

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H., 2021. Transformer tracking, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8122–8131. doi:10.1109/CVPR46437.2021.00803.

Cheng, S., Zhu, Y., Wu, S., 2023. Deep learning based efficient ship detection from drone-captured images for maritime surveillance. Ocean Engineering 285, 115440. URL: https://www.sciencedirect.com/science/article/pii/S0029801823018243, doi:https://doi.org/10.1016/j.oceaneng.2023.115440.

Cui, Y., Jiang, C., Wang, L., Wu, G., 2022. Mixformer: End-to-end tracking with iterative mixed attention, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13598–13608. doi:10.1109/CVPR52688.2022.01324.

Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M., 2017a. Eco: Efficient convolution operators for tracking, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6931–6939. doi:10.1109/CVPR.2017.733.

Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M., 2019. Atom: Accurate tracking by overlap maximization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Danelljan, M., Häger, G., Khan, F.S., Felsberg, M., 2017b. Discriminative scale space tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 1561–1575. doi:10.1109/TPAMI.2016.2609928.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=YicbFdNTTy.

Du, D., Zhu, P., Wen, L., Bian, X., Ling, H., Hu, Q., Zheng, J., Peng, T., Wang, X., Zhang, Y., Bo, L., Shi, H., Zhu, R., Han, B., Zhang, C., Liu, G., Wu, H., Wen, H., Wang, H., Fan, J., Chen, J., Gao, J., Zhang, J., Zhou, J., Zhou, J., Wang, J., Wan, J., Kittler, J., Zhang, K., Huang, K., Yang, K., Zhang, K., Huang, L., Zhou, L., Shi, L., Ding, L., Wang, N., Wang, P., Hu, Q., Laganiere, R., Ma, R., Zhang, R., Zou, S., Zhao, S., Li, S., Zhu, S., Li, S., Ge, S., Xuan, S., Xu, T., 2019. Visdrone-sot2019: The vision meets drone single object tracking challenge results, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops.

Du, G., Zhou, P., Yadikar, N., Aysa, A., Ubul, K., 2025. Mamba meets tracker: exploiting token aggregation and diffusion for robust unmanned aerial vehicles tracking. Complex & Intelligent Systems 11, 204. URL: https://doi.org/10.1007/s40747-025-01821-z, doi:10.1007/s40747-025-01821-z.

Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H., 2019. Lasot: A high-quality benchmark for large-scale single object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y., 2023. Eva: Exploring the limits of masked visual representation learning at scale, in: 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR), IEEE. pp. 19358–19369.

Fu, C., Cao, Z., Li, Y., Ye, J., Feng, C., 2021a. Onboard real-time aerial tracking with efficient siamese anchor proposal network, in: IEEE Transactions on Geoscience and Remote Sensing, pp. 1–13.

Fu, C., Cao, Z., Li, Y., Ye, J., Feng, C., 2021b. Siamese anchor proposal network for high-speed aerial tracking, in: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 510–516. doi:10.1109/ICRA48506.2021.9560756.

Fu, C., Lei, X., Zuo, H., Yao, L., Zheng, G., Pan, J., 2024. Progressive representation learning for real-time uav tracking, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5072–5079. doi:10.1109/IROS58592.2024.10803050.

Galoogahi, H.K., Fagg, A., Lucey, S., 2017. Learning background-aware correlation filters for visual tracking, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1144–1152. doi:10.1109/ICCV.2017.129.

Gopal, G.Y., Amer, M.A., 2024. Separable self and mixed attention transformers for efficient object tracking, in: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 6694–6703. doi:10.1109/WACV57701.2024.00657.

Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2015. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 583–596. doi:10.1109/TPAMI.2014.2345390.

Hu, X., Zhao, J., Hui, Y., Li, S., You, S., 2023. Siamhsft: A siamese network-based tracker with hierarchical sparse fusion and transformer for uav tracking. Sensors 23. URL: https://www.mdpi.com/1424-8220/23/21/8666, doi:10.3390/s23218666.

Huang, L., Zhao, X., Huang, K., 2021. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 1562–1577. doi:10.1109/TPAMI.2019.2957464.

Huang, Z., Fu, C., Li, Y., Lin, F., Lu, P., 2019. Learning aberrance repressed correlation filters for real-time uav tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Kumar, M., Mondal, S., 2021. Recent developments on target tracking problems: A review. Ocean Engineering 236, 109558. URL: https://www.sciencedirect.com/science/article/pii/S0029801821009471, doi:https://doi.org/10.1016/j.oceaneng.2021.109558.

Lan, X., Yang, Z., Zhang, W., Yuen, P.C., 2021. Spatial-temporal regularized multi-modality correlation filters for tracking with re-detection. ACM Trans. Multimedia Comput. Commun. Appl. 17. URL: https://doi.org/10.1145/3430257, doi:10.1145/3430257.

Law, H., Deng, J., 2018. Cornernet: Detecting objects as paired keypoints, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham. pp. 765–781.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J., 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4277–4286. doi:10.1109/CVPR.2019.00441.

Li, J., Zhang, G., Jiang, C., Zhang, W., 2023. A survey of maritime unmanned search system: Theory, applications and future directions. Ocean Engineering 285, 115359. URL: https://www.sciencedirect.com/science/article/pii/S0029801823017432, doi:https://doi.org/10.1016/j.oceaneng.2023.115359.

Li, S., Liu, Y., Zhao, Q., Feng, Z., 2022a. Learning residue-aware correlation filters and refining scale for real-time uav tracking. Pattern Recognition 127, 108614. URL: https://www.sciencedirect.com/science/article/pii/S0031320322000954, doi:https://doi.org/10.1016/j.patcog.2022.108614.

Li, S., Yang, X., Wang, X., Zeng, D., Ye, H., Zhao, Q., 2024a. Learning target-aware vision transformers for real-time uav tracking. IEEE Transactions on Geoscience and Remote Sensing 62, 1–18. doi:10.1109/TGRS.2024.3417400.

Li, S., Yeung, D.Y., 2017. Visual object tracking for unmanned aerial vehicles: a benchmark and new motion models, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press. p. 4140–4146.

Li, Y., Fu, C., Ding, F., Huang, Z., Lu, G., 2020. Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11920–11929. doi:10.1109/CVPR42600.2020.01194.

Li, Y., Liu, M., Wu, Y., Wang, X., Yang, X., Li, S., 2024b. Learning adaptive and view-invariant vision transformer for real-time UAV tracking, in: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (Eds.), Proceedings of the 41st International Conference on Machine Learning, PMLR. pp. 28403–28420. URL: https://proceedings.mlr.press/v235/li24ax.html.

Li, Y., Mao, H., Girshick, R., He, K., 2022b. Exploring plain vision transformer backbones for object detection, in: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham. pp. 280–296.

Li, Y., Wang, B., Wu, X., Liu, Z., Li, Y., 2024c. Lightweight full-convolutional siamese tracker. Knowledge-Based Systems 286, 111439. URL: https://www.sciencedirect.com/science/article/pii/S0950705124000741, doi:https://doi.org/10.1016/j.knosys.2024.111439.

Lin, B., Zheng, J., Xue, C., Fu, L., Li, Y., Shen, Q., 2024. Motion-aware correlation filter-based object tracking in satellite videos. IEEE Transactions on Geoscience and Remote Sensing 62, 1–13.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944. doi:10.1109/CVPR.2017.106.

Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European Conference on Computer Vision. URL: https://api.semanticscholar.org/CorpusID:14113767.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8759–8768. doi:10.1109/CVPR.2018.00913.

Liu, X., Xiao, J., Dong, X., 2025a. Underwater single target tracking with self-prompting. Intelligent Marine Technology and Systems 3, 1–13.

Liu, X., Zhou, H., Dong, J., Xiao, J., Dong, X., 2025b. Single target tracking with contextual prompts in scene understanding. IEEE Transactions on Artificial Intelligence PP.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization .

Luo, C., Yang, X., Yuille, A., 2021. Exploring simple 3d multi-object tracking for autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10488–10497.

Mueller, M., Smith, N., Ghanem, B., 2016. A benchmark and simulator for uav tracking, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 445–461.

Müller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B., 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham. pp. 310–327.

Novák, F., Báča, T., Procházka, O., Saska, M., 2025. State estimation of marine vessels affected by waves by unmanned aerial vehicles. Ocean Engineering 323, 120606. URL: https://www.sciencedirect.com/science/article/pii/S002980182500321X, doi:https://doi.org/10.1016/j.oceaneng.2025.120606.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666. doi:10.1109/CVPR.2019.00075.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H., 2021. Training data-efficient image transformers amp; distillation through attention, in: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 10347–10357. URL: https://proceedings.mlr.press/v139/touvron21a.html.

Varga, L.A., Kiefer, B., Messmer, M., Zell, A., 2022. Seadronessee: A maritime benchmark for detecting humans in open water, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2260–2270.

Wang, K., Wang, Z., Zhang, X., Liu, M., 2025. Bstrack: Robust uav tracking using feature extraction of bilateral filters and sparse attention mechanism. Expert Systems with Applications 267, 126202. URL: https://www.sciencedirect.com/science/article/pii/S0957417424030690, doi:https://doi.org/10.1016/j.eswa.2024.126202.

Wang, N., Zhou, W., Tian, Q., Hong, R., Wang, M., Li, H., 2018. Multi-cue correlation filters for robust visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, Y., Luo, J., Jia, W., 2026. Lightweight network for small target detection and multi-target tracking in unmanned aerial vehicle maritime search and rescue, in: 2026 6th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS), pp. 202–206. doi:10.23919/ACCTCSCPS00012.2026.00042.

Wen, S., Meriaux, E., Guzmán, M.S., Wang, Z., Shi, J., Dudek, G., 2026. Stable multi-drone gnss tracking system for marine robots, in: 2026 IEEE International Conference on Robotics and Automation (ICRA), pp. 11965–11972. doi:10.1109/ICRA57385.2026.11695857.

Wu, Y., Li, Y., Liu, M., Wang, X., Yang, X., Ye, H., Zeng, D., Zhao, Q., Li, S., 2025a. Learning an adaptive and view-invariant vision transformer for real-time uav tracking. IEEE Transactions on Circuits and Systems for Video Technology , 1–1doi:10.1109/TCSVT.2025.3599856.

Wu, Y., Wang, X., Yang, X., Liu, M., Zeng, D., Ye, H., Li, S., 2025b. Learning occlusion-robust vision transformers for real-time uav tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17103–17113.

Xie, F., Yang, W., Wang, C., Chu, L., Cao, Y., Ma, C., Zeng, W., 2024. Correlation-embedded transformer tracking: A single-branch framework. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 10681–10696.

Xue, C., Liang, Q., Zhong, B., Zu, Y., Xue, Y., Xia, H., Song, S., 2026. Toward low-cost yet effective temporal learning for uav tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 42538–42548.

Xue, C., Zhong, B., Liang, Q., Xia, H., Song, S., 2024. Unifying motion and appearance cues for visual tracking via shared queries. IEEE Transactions on Circuits and Systems for Video Technology 35, 1987–1998.

Xue, C., Zhong, B., Liang, Q., Zheng, Y., Li, N., Xue, Y., Song, S., 2025. Similarity-guided layer-adaptive vision transformer for uav tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6730–6740.

Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021a. Learning spatio-temporal transformer for visual tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10448–10457.

Yan, B., Peng, H., Wu, K., Wang, D., Fu, J., Lu, H., 2021b. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15180–15189.

Yao, J., Huang, N., Shi, W., Han, X., Yu, C., 2026. Dgp-sfnet: An efficient visual perception architecture for small maritime objects on unmanned surface vehicles. Ocean Engineering 363, 126533.

Yao, L., Fu, C., Li, S., Zheng, G., Ye, J., 2023. Sgdvit: Saliency-guided dynamic vision transformer for uav tracking, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 3353–3359. doi:10.1109/ICRA48891.2023.10161487.

Ye, B., Chang, H., Ma, B., Shan, S., Chen, X., 2022. Joint feature learning and relation modeling for tracking: A one-stream framework, in: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham. pp. 341–357.

Yu, C., Wang, L., Sun, G., Guo, S., Liang, X., 2026. Geometry-guided and interaction-aware multi-agent reinforcement learning for cooperative multi-usv encirclement. Ocean Engineering 367, 127899.

Yu, C., Zhang, Z., Liu, X., Yang, C., Wang, Y., Liang, X., 2025. Verification method for formation control of usvs based on virtual-real integration. Ocean Engineering 335, 121677.

Zhang, Z., Peng, H., Fu, J., Li, B., Hu, W., 2020. Ocean: Object-aware anchor-free tracking, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham. pp. 771–787.

Zhao, Z., Shen, H., Liang, X., Wang, L., Han, B., 2025. Construction of a virtual dataset of maritime search and rescue targets for unmanned aerial vehicles. Ocean Engineering 328, 120926. URL: https://www.sciencedirect.com/science/article/pii/S0029801825006390, doi:https://doi.org/10.1016/j.oceaneng.2025.120926.

Zu, Y., Liang, Q., Xue, C., Tan, Y., Song, S., Zhong, B., 2026. Robust nighttime uav tracking via parameter and computation efficient transfer learning. IEEE Transactions on Multimedia .

Zuo, H., Fu, C., Li, S., Lu, K., Li, Y., Feng, C., 2023. Adversarial blur-deblur network for robust uav tracking. IEEE Robotics and Automation Letters 8, 1101–1108. doi:10.1109/LRA.2023.3236584.