When Trackers Date Fish: A Benchmark and Framework for Underwater Multiple Fish Tracking

Weiran Li[1,2], Yeqiang Liu[1], Qiannan Guo[1], Yijie Wei[1], Hwa Liang Leo[2], Zhenbo Li[1]*

[1]China Agricultural University
[2]National University of Singapore
{vranlee,yeqiangliu,guoqiannan,yjwei,lizb}@cau.edu.cn
weiranli@u.nus.edu, bielhl@nus.edu.sg

Abstract

Multiple object tracking (MOT) technology has made significant progress in terrestrial applications, but underwater tracking scenarios remain underexplored despite their importance to marine ecology and aquaculture. In this paper, we present *Multiple Fish Tracking Dataset 2025* (MFT25), a comprehensive dataset specifically designed for underwater multiple fish tracking, featuring 15 diverse video sequences with 408,578 meticulously annotated bounding boxes across 48,066 frames. Our dataset captures various underwater environments, fish species, and challenging conditions including occlusions, similar appearances, and erratic motion patterns. Additionally, we introduce *Scale-aware and Unscented Tracker* (SU-T), a specialized tracking framework featuring an *Unscented Kalman Filter* (UKF) optimized for non-linear swimming patterns of fish and a novel *Fish-Intersection-over-Union* (FishIoU) matching that accounts for the unique morphological characteristics of aquatic species. Extensive experiments demonstrate that our SU-T baseline achieves state-of-the-art performance on MFT25, with 34.1 HOTA and 44.6 IDF1, while revealing fundamental differences between fish tracking and terrestrial object tracking scenarios.

Codes and Dataset — https://vranlee.github.io/SU-T
Extended version — https://arxiv.org/abs/2507.06400

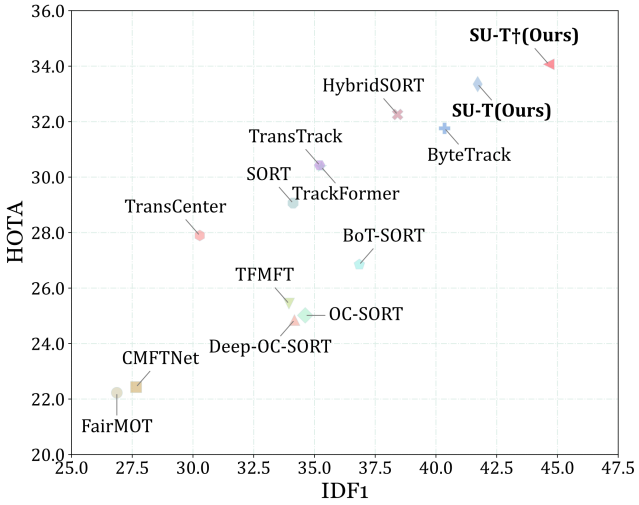

Figure 1: Evaluation of various MFT and MOT methods on MFT25 benchmark. Detailed results are provided in Table 2.

Introduction

A tracker's greatest challenge is not merely to find a fish, but to arrange a ***date*** *with the same fleeting shadow—****a perfect alignment of time and space****.*

(Preface)

Fish behavior monitoring and group dynamics analysis form essential technical foundations for marine ecological research, aquaculture optimization, and fishery resource management (Cui et al. 2024; Huang et al. 2018). With advancements in computer vision and deep learning, underwater Multiple Fish Tracking (MFT) has emerged as a core technology for efficient, non-invasive observation (Zeng et al. 2023; Li et al. 2018). It enables quantitative analysis of fish movement patterns, group interactions, and environmental adaptation mechanisms by continuously tracking and associating individual targets across video sequences. MFT offers significant applications in endangered species protection, aquaculture density optimization, and marine ecosystem modeling (Jager et al. 2017).

MFT is a specialized application of Multiple Object Tracking (MOT), presents unique challenges in underwater environments (Hassan et al. 2024; Dendorfer et al. 2021). It aims to generate continuous trajectories of individual fish through reliable identification across video frames. Unlike single fish tracking, MFT must distinguish between numerous similar-looking fish and maintain consistent identity assignments despite rapid direction changes and frequent occlusions (Bewley et al. 2016; Zhang et al. 2022). The ability to resolve confusion between morphologically similar individuals in varying water conditions becomes critical to successful fish tracking, particularly in dense shoaling scenarios where individuals frequently cross paths (Sun et al. 2022).

Current methods largely rely on data-driven approaches, leveraging high-precision detectors to obtain real-time target positions (Zhang et al. 2021; Li, Li, and Li 2022). However, tracking fish in complex underwater environments presents

*Corresponding author

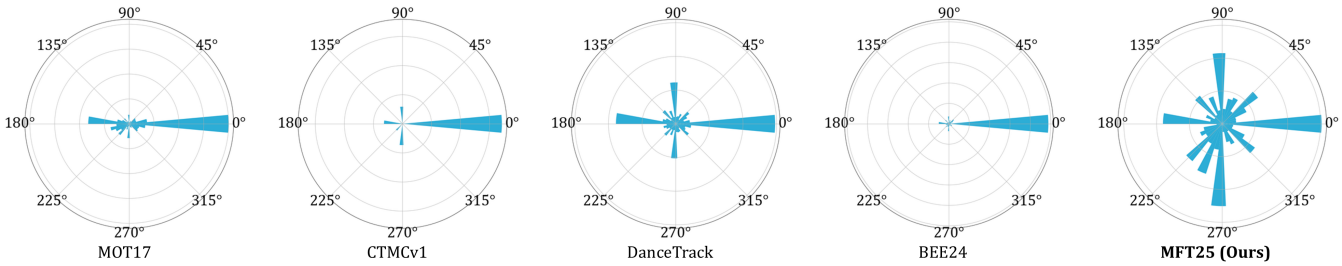

Figure 2: Distribution of target movement directions across datasets. Directional instability is notably more pronounced in the fish dataset compared to other target categories.

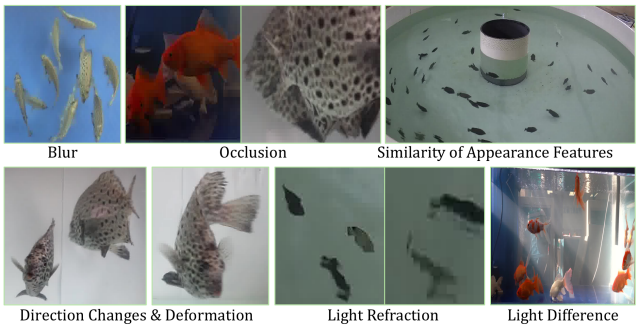

Figure 3: The challenges of multiple fish tracking arise from factors such as fish physiological features and the complexity of underwater scenarios.

several challenges, as shown in Fig. 3. On the one hand, high morphological similarity among individual fish combined with their erratic movement patterns frequently leads to identity switches and trajectory fragmentation (Li et al. 2024b). On the other hand, existing public datasets suffer from insufficient diversity and poor image quality (Pedersen et al. 2023, 2020), limiting the development of tracking models with strong generalization capabilities across complex scenarios.

To address these challenges, we present *Multiple Fish Tracking Dataset 2025* (MFT25), a large-scale dataset specifically designed for underwater MOT task, alongside *Scale-aware and Unscented Tracker* (SU-T), an efficient, lightweight baseline model for online tracking. Our dataset and tracker aim to establish a robust foundation for advancing research in underwater object tracking systems with practical applications in marine ecology and aquaculture. Our main contributions are summarized as follows:

- We introduce MFT25, a large-scale fish dataset for MOT, featuring 15 diverse video sequences with 408,578 meticulously annotated bounding boxes across 48,066 frames, capturing various underwater environments, fish species, and challenging conditions including occlusions, rapid direction changes, and visually similar appearances.
- We propose SU-T, a specialized tracking framework featuring an *Unscented Kalman Filter* (UKF) optimized for non-linear fish swimming patterns and a novel *Fish-Intersection-over-Union* (FishIoU) matching that accounts for the unique morphological characteristics and erratic movement behaviors of aquatic species.
- We conduct extensive comparative experiments demonstrating that our tracker achieves state-of-the-art performance on MFT25, with 34.1 HOTA and 44.6 IDF1, as illustrated in Fig. 1.
- Through quantitative analysis, we highlight the fundamental differences between fish tracking and land-based object tracking scenarios, as shown in Fig. 2 and *extended version*.

MFT25: When Trackers Date Fish

Multiple Fish Tracking

Related Methods Fish tracking presents unique challenges due to complex underwater environments, distinctive morphology, and erratic swimming behaviors (Cui et al. 2024). Early approaches used traditional techniques like background subtraction (Shevchenko, Eerola, and Kaarna 2018) and object segmentation (Huang et al. 2018), while recent advances employ deep learning methods including SiamRPN (Wang et al. 2022), appearance-based models (Li et al. 2018), graph-based tracking (Jager et al. 2017), and Swin Transformers (Zeng et al. 2023). While sharing principles with terrestrial tracking, underwater MOT remains less developed (Hassan et al. 2024). Terrestrial MOT research has explored multi-modal fusion (Li et al. 2024a), adaptive frame rates (Liu, Wu, and Fu 2023), computational efficiency (Liu, Li, and Wang 2023), appearance modeling (Seidenschwarz et al. 2023), depth integration (Liu et al. 2025), and high-density scenarios (Lei et al. 2024).

Contemporary frameworks fall into three categories: Joint-Detection-Embedding (JDE) methods use unified networks achieving training efficiency but compromising specialization (Zhang et al. 2021; Li, Li, and Li 2022); Transformer approaches leverage attention mechanisms with high performance but substantial overhead (Dosovitskiy et al. 2020; Li et al. 2024b); Separated-Detection-Embedding (SDE) methods extend SORT (Bewley et al. 2016), decoupling detectors from appearance models for balanced accuracy and efficiency (Zhang et al. 2022; Cao et al. 2023; Xiao et al. 2024; Fischer et al. 2023; Aharon, Orfaig, and Bobrovsky 2022; Yang et al. 2024). Recent innovations include camera pose estimation (Yi et al. 2024), generative diffusion

models (Luo et al. 2024), and pixel-wise trajectory propagation (Zhao et al. 2022), but show limited underwater adaptability and insufficient real-time performance for practical fish tracking applications.

Related Datasets MOT datasets require frame-by-frame bounding box annotations with consistent identity information across extended video sequences, representing a significant annotation challenge (Dendorfer et al. 2020). Current MOT research predominantly focuses on terrestrial domains, resulting in well-established benchmarks for humans, vehicles, and animals, such as the MOT challenge series (Dendorfer et al. 2021), DanceTrack (Sun et al. 2022), BEE24 (Cao et al. 2025), and CTMC (Anjum and Gurari 2020).

While several fish-oriented video datasets exist, including Fish4Knowledge, DeepFish, and SeaCLEF (Cui et al. 2024), they present significant limitations for modern tracking applications. These early datasets typically suffer from low resolution, poor visibility conditions that obscure fish identities, and inconsistent annotation formats that impede effective model training. Furthermore, other specialized datasets, such as FishTrack23 (Dawkins et al. 2024) and WebUOT-1M (Zhang et al. 2024), are designed for Single Object Tracking (SOT), a fundamentally different task that provides annotations only for the initialized target. In contrast, MOT datasets require frame-by-frame annotations for all visible targets, along with consistent identity labels to enable long-term association. The CFC dataset (Kay et al. 2022) utilizes sonar imaging, which represents a distinct data modality from the optical videos commonly used in aquaculture applications. More recent standardized fish tracking datasets, including BrackishMOT (Pedersen et al. 2023), 3D-ZeF (Pedersen et al. 2020), and MFT22 (Li et al. 2024b), have emerged with consistent annotation protocols. However, these datasets remain limited in both environmental diversity and scale, typically featuring simplified scenarios under controlled conditions. The absence of comprehensive, high-quality, and standardized fish tracking datasets thus represents a critical bottleneck that constrains significant advances in underwater MFT research.

Dataset Construction

The MFT25 dataset was captured using imaging equipment of Canon $EOSR6$ and Sony $\alpha 7M3$, across diverse aquaculture environments. Recording locations encompassed both industrial circulating water aquaculture ponds and controlled laboratory tanks to ensure environmental diversity. The dataset features multiple fish species with distinctly different morphologies, including commercially valuable groupers and ornamental koi at various developmental stages, introducing substantial appearance variation.

To ensure comprehensive scenario coverage, we systematically deployed multiple camera configurations, including both overhead and horizontal perspectives, across varied illumination conditions from daylight to nocturnal settings. The dataset consists exclusively of authentic footage without synthetic augmentation, preserving the natural complexity of scenarios. All bounding box annotations were created using DarkLabel software through manual selection and verification processes. The resulting MFT25 dataset encompasses 15 diverse video sequences containing 223 distinct fish trajectories across 48,066 frames, with a total of 408,578 precisely annotated bounding boxes. This represents a substantial advancement in scale, containing 2.6-8.3 times more annotated instances compared to previous fish tracking datasets. Table 1 presents a comprehensive statistical comparison with existing fish tracking benchmarks.

Dataset	BrackishMOT	3D-ZeF	MFT22	MFT25 (Ours)
Clips	98	8	10	15
Tracks	638	32	234	223
FPS	25	60	25	25
Frames	14,017	14,398	9,100	**48,066**
Boxes	49,364	86,452	155,437	**408,578**

Table 1: Quantitative comparison of MFT datasets.

Moreover, we analyze the distinctive movement characteristics of fish in our dataset through detailed quantitative analysis in *extended version*.

SU-T: A MFT Baseline

Framework

To address the unique challenges of MFT, we propose *Scale-aware and Unscented Tracker* (SU-T), a specialized baseline following the SDE paradigm. As illustrated in Fig. 4, our framework comprises three primary components: a detector, an association module, and an optional Re-Identification (Re-ID) module. The processing pipeline begins with video frames being fed into the detector, which generates bounding boxes with corresponding confidence scores. These detections are then processed by the association module, where our specialized FishIoU metric calculates matching costs between detected boxes and predictions from the UKF. The Hungarian algorithm performs optimal assignment to update existing trajectories and establish new tracks when necessary. To address the challenge of visually similar fish, SU-T integrates an optional Re-ID module that extracts discriminative feature embeddings. These embeddings work synergistically with the FishIoU metric during association, significantly enhancing tracking accuracy by maintaining consistent identities even when fish exhibit nearly identical appearances.

Detector and Re-ID

Considering the variability of fish movements and the significant scale variations due to varying distances from the camera, our tracker adopts a mainstream pyramid-based design following (Zhang et al. 2022; Cao et al. 2023; Yang et al. 2024) and employs decoupled heads to predict point centers, bounding boxes, offsets, and confidence scores (Ge et al. 2021). The different decoupled heads share feature parameters across pyramid levels. Additionally, an optional re-identification module continuously updates GeM Pooling through learnable parameters, outputting target appearance

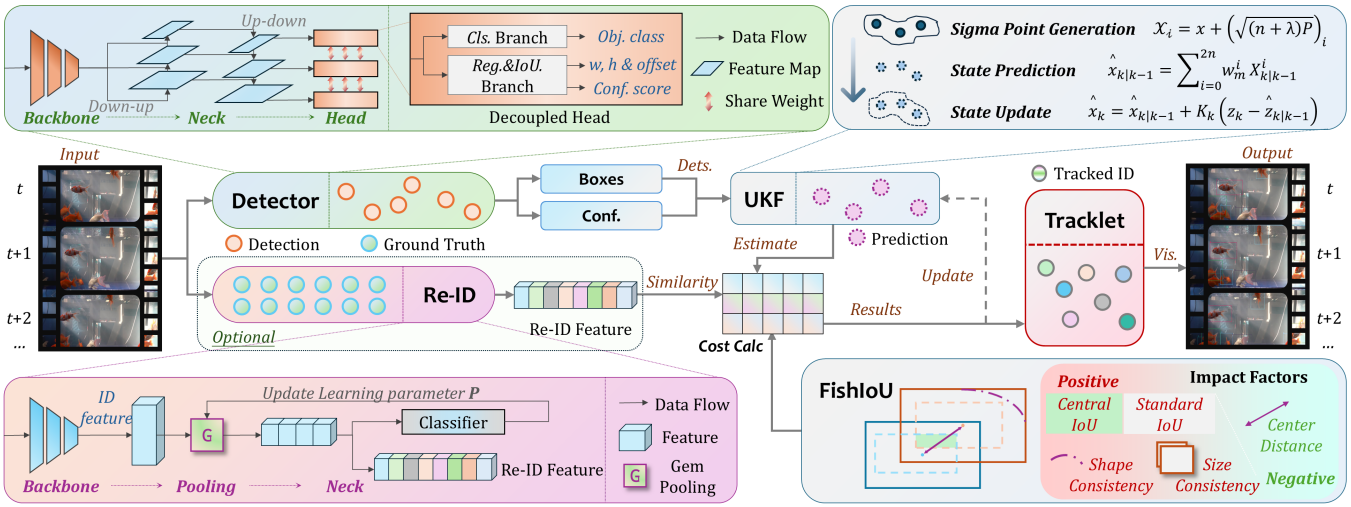

Figure 4: The framework of SU-T. The pipeline consists of three main components: a detector, an association module, and an optional Re-ID module. The detector generates bounding boxes and confidence scores for each frame. The optional Re-ID module extracts feature embeddings to enhance tracking accuracy. The association module uses the FishIoU to calculate matching costs between detected boxes and predicted boxes from the UKF.

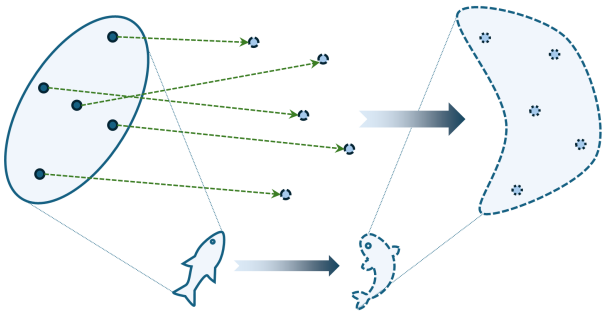

Figure 5: *Unscented Kalman Filter* (UKF) motion model used in SU-T for predicting fish movement in complex underwater environments.

features that are incorporated into the cost matrix for subsequent association, thereby providing robust appearance similarity cues for tracking.

Unscented Kalman Filter

The *Unscented Kalman Filter* (UKF) is particularly well-suited for tracking fish due to their non-linear motion patterns, as shown in Fig. 5. Unlike standard Kalman Filter, UKF uses a deterministic sampling technique to handle non-linearities. The three core mathematical components of our UKF implementation are as follows.

Sigma Points Generation For state vector $\mathbf{x} \in \mathbb{R}^n$ with covariance $\mathbf{P}$, we generate $2n+1$ sigma points:

$$\mathcal{X}_0 = \mathbf{x} \tag{1}$$

$$\mathcal{X}_i = \mathbf{x} + \left(\sqrt{(n+\lambda)\mathbf{P}}\right)_i, \quad i = 1, \ldots, n \tag{2}$$

$$\mathcal{X}_{i+n} = \mathbf{x} - \left(\sqrt{(n+\lambda)\mathbf{P}}\right)_i, \quad i = 1, \ldots, n \tag{3}$$

where $\lambda = \alpha^2(n+\kappa) - n$ is a scaling parameter, α controls spread of points, κ is a secondary parameter (typically 3-n), and $\left(\sqrt{(n+\lambda)\mathbf{P}}\right)_i$ is the i-th column of the matrix square root.

Prediction Step Each sigma point is propagated through the non-linear state transition function $\mathbf{f}$ at time step k:

$$\mathcal{X}^i_{k|k-1} = \mathbf{f}(\mathcal{X}^i_{k-1}), \quad i = 0, \ldots, 2n \tag{4}$$

$$\hat{\mathbf{x}}_{k|k-1} = \sum_{i=0}^{2n} w^i_m \mathcal{X}^i_{k|k-1} \tag{5}$$

$$\mathbf{P}_{k|k-1} = \sum_{i=0}^{2n} w^i_c [\mathcal{X}^i_{k|k-1} - \hat{\mathbf{x}}_{k|k-1}][\mathcal{X}^i_{k|k-1} - \hat{\mathbf{x}}_{k|k-1}]^T + \mathbf{Q}_k \tag{6}$$

where w^i_m and w^i_c are weight coefficients for mean and covariance, and $\mathbf{Q}_k$ is the process noise covariance.

Measurement Update Step The predicted sigma points are transformed through the measurement function $\mathbf{h}$:

$$\mathcal{Z}^i_{k|k-1} = \mathbf{h}(\mathcal{X}^i_{k|k-1}), \quad i = 0, \ldots, 2n \tag{7}$$

$$\hat{\mathbf{z}}_{k|k-1} = \sum_{i=0}^{2n} w^i_m \mathcal{Z}^i_{k|k-1} \tag{8}$$

$$\mathbf{P}_{zz} = \sum_{i=0}^{2n} w^i_c [\mathcal{Z}^i_{k|k-1} - \hat{\mathbf{z}}_{k|k-1}][\mathcal{Z}^i_{k|k-1} - \hat{\mathbf{z}}_{k|k-1}]^T + \mathbf{R}_k \tag{9}$$

$$\mathbf{P}_{xz} = \sum_{i=0}^{2n} w_c^i [\mathcal{X}_{k|k-1}^i - \hat{\mathbf{x}}_{k|k-1}][\mathcal{Z}_{k|k-1}^i - \hat{\mathbf{z}}_{k|k-1}]^T \quad (10)$$

$$\mathbf{K}_k = \mathbf{P}_{xz}\mathbf{P}_{zz}^{-1} \quad (11)$$

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{z}_k - \hat{\mathbf{z}}_{k|k-1}) \quad (12)$$

$$\mathbf{P}_k = \mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{P}_{zz}\mathbf{K}_k^T \quad (13)$$

where $\mathbf{z}_k$ is the actual measurement, $\mathbf{R}_k$ is the measurement noise covariance, $\mathbf{P}_{zz}$ and $\mathbf{P}_{xz}$ are the measurement and cross-covariances, and $\mathbf{K}_k$ is the Kalman gain.

FishIoU: Scale-aware Association

The unique morphology and movement patterns of fish present significant challenges for standard object association. We introduce *Fish-Intersection-over-Union* (FishIoU), a specialized association IoU that accounts for the elongated body structure, erratic motion patterns, and size variations common in fish species.

Given two bounding boxes $B_1 = [x_1, y_1, x_2, y_2]$ and $B_2 = [x_1', y_1', x_2', y_2']$, we first compute the standard IoU as:

$$\text{IoU} = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|} \quad (14)$$

where $|B_1 \cap B_2|$ represents the intersection area and $|B_1 \cup B_2|$ the union area. Then, to account for fish morphology, we incorporate a center distance penalty:

$$d_c = \frac{(c_x - c_x')^2 + (c_y - c_y')^2}{d_{\text{diag}}^2} \quad (15)$$

where (c_x, c_y) and (c_x', c_y') are the centers of B_1 and B_2 respectively, and d_{diag}^2 is the squared diagonal length of the enclosing box. Considering fish bodies typically have an elongated structure with important features concentrated in the front, we define a central region for each box with asymmetric insets to emphasize this characteristic:

$$B_1^c = [x_1 + \alpha w_1,\ y_1 + \beta h_1,\ x_2 - \gamma w_1,\ y_2 - \beta h_1] \quad (16)$$

$$B_2^c = [x_1' + \alpha w_2,\ y_1' + \beta h_2,\ x_2' - \gamma w_2,\ y_2' - \beta h_2] \quad (17)$$

where w_1, h_1 and w_2, h_2 are the width and height of B_1 and B_2 respectively, and α, β, γ are constant factors determined empirically based on fish morphological characteristics. By default, $\alpha = 0.15$, $\beta = 0.3$, and $\gamma = 0.25$. The central IoU is then calculated as:

$$\text{cIoU} = \frac{|B_1^c \cap B_2^c|}{|B_1^c \cup B_2^c|} \quad (18)$$

To account for consistent fish orientation, we consider the aspect ratio consistency:

$$\alpha_r = \frac{\min(r_1, r_2)}{\max(r_1, r_2)} \quad (19)$$

where $r_1 = \frac{w_1}{h_1}$ and $r_2 = \frac{w_2}{h_2}$ are the aspect ratios of the two boxes.

Additionally, we incorporate area ratio consistency, since the size of the fish does not change abruptly between frames:

$$\alpha_a = \frac{\min(a_1, a_2)}{\max(a_1, a_2)} \quad (20)$$

where $a_1 = w_1 \times h_1$ and $a_2 = w_2 \times h_2$ are the areas of the two boxes.

For small targets, we apply a scale factor to reduce the center distance penalty:

$$s = 1 - e^{-\frac{\min(a_1, a_2)}{1000}} \quad (21)$$

The final FishIoU metric combines these components with specific weights optimized for fish tracking:

$$\text{FishIoU} = \omega_1 \cdot \text{IoU} + \omega_2 \cdot \text{cIoU} + \omega_3 \cdot \alpha_r + \omega_4 \cdot \alpha_a - \omega_5 \cdot s \cdot d_c \quad (22)$$

where $\omega_1 = 1.0$, $\omega_2 = 0.3$, $\omega_3 = 0.1$, $\omega_4 = 0.2$, and $\omega_5 = 0.4$ are weights determined empirically through extensive experiments.

Assocation

Our association strategy employs progressive confidence-based processing that significantly reduces identity switches while maintaining computational efficiency. Algorithm 1 presents the complete multi-level cascade tracking process, which integrates our UKF motion prediction and FishIoU matching with a cascaded association strategy for aquatic environments.

Building upon frameworks from HybridSORT (Yang et al. 2024), we adopt dual-confidence matching of ByteTrack (Zhang et al. 2022) and heuristic observation-centric recovery of OC-SORT (Cao et al. 2023), and implements three association stages. In the first stage, high-confidence detections are matched with existing tracks using our specialized FishIoU, establishing reliable primary associations even when fish exhibit rapid direction changes. The second stage associates remaining tracks with low-confidence detections, effectively recovering temporarily occluded targets while filtering false positives induced by water turbidity and reflections. The final stage attempts to reconnect tracks with their historical appearances, addressing the frequent, abrupt directional changes and non-linear swimming patterns characteristic of fish locomotion.

For cost calculation, our framework integrates both spatial and appearance information when available. Spatially, the FishIoU outperforms standard IoU by incorporating fish-specific morphological features into the matching process. When enabled, the Re-ID module provides discriminative appearance embeddings that effectively differentiate between visually similar individuals swimming in close proximity.

Experiments

Implementation Details

For all experiments, we employed the same YOLOX (Ge et al. 2021) detector and the same Re-ID network (He

Algorithm 1: Multi-Level Cascade Tracking

Input: Detections $\mathcal{D}$ with scores, Existing tracks $\mathcal{T}$
Output: Updated tracks $\mathcal{T}$
/* Prediction */
for $T_j \in \mathcal{T}$ **do**
  $\hat{B}_j, s_j \leftarrow$ UKF.predict(T_j) /* Predict box and score */
end
/* First Association: High-confidence */
$\mathcal{D}_{high} \leftarrow \{d_i \in \mathcal{D} : score(d_i) > \tau_{high}\}$
$\mathbf{C} \leftarrow$ FishIoU($\mathcal{D}_{high}, \{\hat{B}_j\}$) /* Base cost */
if *Use Re-ID* **then**
  $\mathbf{C} \leftarrow \omega_1 \mathbf{C} + \omega_2$ EmbeddingDistance($\mathcal{D}_{high}, \mathcal{T}$)
end
$\mathcal{M}_1, \mathcal{U}_\mathcal{D}, \mathcal{U}_\mathcal{T} \leftarrow$ Hungarian($-\mathbf{C}$)
for $(i, j) \in \mathcal{M}_1$ **do**
  T_j.update(d_i)
end
/* Second Association: Low-confidence */
$\mathcal{D}_{low} \leftarrow \{d_i \in \mathcal{D} : \tau_{low} < score(d_i) < \tau_{high}\}$
$\mathbf{C}_{iou} \leftarrow$ FishIoU($\mathcal{D}_{low}, \{\hat{B}_j : T_j \in \mathcal{U}_\mathcal{T}\}$)
if *Use Re-ID* **then**
  $\mathbf{C}_{iou} \leftarrow \mathbf{C}_{iou} + \lambda \cdot$ EmbeddingDistance($\mathcal{D}_{low}$, $\{T_j \in \mathcal{U}_\mathcal{T}\}$)
end
$\mathcal{M}_2 \leftarrow$ Hungarian($-\mathbf{C}_{iou}$)
for $(i, j) \in \mathcal{M}_2$ *where* $\mathbf{C}_{iou}[i, j] > \tau_{iou}$ **do**
  T_j.update($\mathcal{D}_{low}[i]$) /* Update state */
  $\mathcal{U}_\mathcal{T} \leftarrow \mathcal{U}_\mathcal{T} \setminus \{T_j\}$
end
/* Third Association: Last-chance */
$\mathbf{C}_{last} \leftarrow$ FishIoU($\mathcal{D}_{high}[\mathcal{U}_\mathcal{D}]$, $\{T_j.last\_observation : T_j \in \mathcal{U}_\mathcal{T}\}$)
for $(i, j) \in$ *Hungarian*$(-\mathbf{C}_{last})$ *where* $\mathbf{C}_{last}[i, j] > \tau_{iou}$ **do**
  T_j.update($\mathcal{D}_{high}[\mathcal{U}_\mathcal{D}[i]]$)
  Remove i from $\mathcal{U}_\mathcal{D}$, T_j from $\mathcal{U}_\mathcal{T}$
end
/* Finalize */
Update all $T_j \in \mathcal{U}_\mathcal{T}$ without observation
Initialize new tracks from $\mathcal{D}_{high}[\mathcal{U}_\mathcal{D}]$
Remove tracks with $time\_since\_update > max\_age$
return $\mathcal{T}$

et al. 2023) for SDE-based models, trained with consistent hyperparameter configurations following the established protocols from ByteTrack (Zhang et al. 2022) and BoT-SORT (Aharon, Orfaig, and Bobrovsky 2022). All models were trained on the MFT25 training set using an NVIDIA A100 GPU, with performance evaluated on the test set using standard MOT metrics, including the comprehensive HOTA (Luiten et al. 2021) metric alongside traditional CLEAR (Bernardin and Stiefelhagen 2008) metrics such as MOTA, IDF1, and ID switches (IDs). Additional experiments, details and discussions are provided in the *extended version*.

Benchmark Results

We compare our proposed SU-T baseline with state-of-the-art MOT and MFT methods on the MFT25 dataset. Table 2 presents the comprehensive comparison results. Our method achieves the best overall performance with 34.1 HOTA. The superiority of our method is particularly evident in association metrics, where SU-T achieves the highest IDF1 score of 44.6 with Re-ID module, significantly outperforming other methods. This demonstrates that our approach better preserves fish identities across frames, enabling more accurate trajectory analysis in challenging underwater scenarios.

Although Transformer-based TrackFormer (Meinhardt et al. 2022) achieves the highest MOTA score of 74.6. However, it exhibits relatively weaker performance in identity preservation metrics such as IDF1 and AssA. Besides, transformer-based trackers impose substantial computational overhead, rendering them impractical for real-time underwater monitoring applications. The performance gap between conventional terrestrial-focused trackers and SU-T validates our hypothesis that underwater tracking scenarios necessitate domain-specific adaptations. Visual comparisons of various tracking methods on MFT25 are illustrated in Fig. 6.

In addition, we conducted additional experiments to evaluate the generalization capability of SU-T on mainstream land-based tracking benchmarks MOT17 and MOT20 (Dendorfer et al. 2021). Our baseline achieved 60.4 and 56.5 HOTA, respectively.

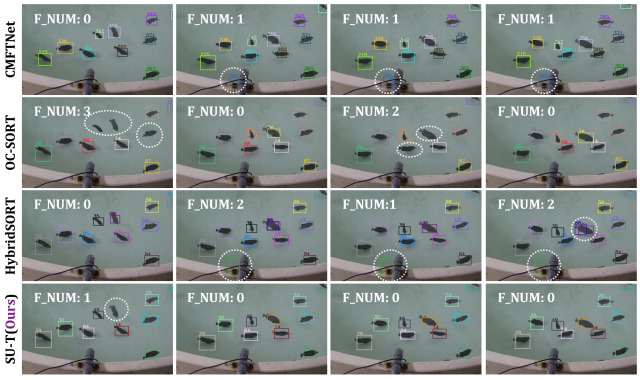

Figure 6: Tracking performance of various trackers on the MFT25 dataset. F_NUM denotes false tracked number, including IDFN, IDFP, and IDs. Best viewed in color.

Ablation Studies

We conducted extensive ablation experiments to evaluate each component. Table 3 reports the influence of different Re-ID models on tracking performance. Among all candidates, the SBS-S101 (He et al. 2023) backbone delivers the best overall results with a HOTA score of 33.8. Interestingly, the IBN-based variants (Pan et al. 2018) do not yield consistent gains, indicating that domain adaptation strategies tailored for terrestrial settings may not transfer well to underwater conditions. Therefore, we adopt SBS-S101 as our final Re-ID module, which offers a strong balance of accuracy,

Method	Class	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	IDs↓	IDFP↓	IDFN↓	Frag↓
FairMOT (Zhang et al. 2021)	JDE	22.226	26.867	47.509	13.910	35.606	939	58198	113393	3768
CMFTNet (Li, Li, and Li 2022)	JDE	22.432	27.659	46.365	14.278	35.452	1301	64754	111263	2769
Deep-OC-SORT (Maggiolino et al. 2023)	SDE	24.848	34.176	46.721	17.537	35.373	550	53478	104024	3659
OC-SORT (Cao et al. 2023)	SDE	25.017	34.620	46.706	17.783	35.369	550	**52934**	103495	3651
TFMFT (Li et al. 2024b)	TransF	25.440	33.950	49.725	17.112	38.059	719	63125	102378	3251
BoT-SORT (Aharon, Orfaig, and Bobrovsky 2022)	SDE	26.848	36.847	49.108	19.446	37.241	500	57581	99181	2704
TransCenter (Xu et al. 2022)	TransF	27.896	30.278	68.693	**30.255**	30.301	807	101223	101002	1992
SORT (Bewley et al. 2016)	SDE	29.063	34.119	69.038	16.952	50.195	778	88928	96815	1726
TrackFormer (Meinhardt et al. 2022)	TransF	30.361	35.285	**74.609**	17.661	**52.649**	718	89391	94720	1729
TransTrack (Sun et al. 2020)	TransF	30.426	35.215	68.983	18.525	50.458	1116	96045	93418	2588
ByteTrack (Zhang et al. 2022)	SDE	31.758	40.355	69.586	20.392	49.712	**489**	80765	87866	**1555**
HybridSORT (Yang et al. 2024)	SDE	32.258	38.421	68.905	20.936	49.992	613	85924	90022	1931
HybridSORT† (Yang et al. 2024)	SDE	32.705	41.727	69.167	21.701	49.697	562	79189	85830	1963
SU-T (Ours)	SDE	33.351	41.717	68.450	22.425	49.943	607	83111	84814	2006
SU-T† (Ours)	SDE	**34.067**	**44.643**	68.958	23.594	49.531	544	76440	**81304**	2011

Table 2: Comparison of different tracking methods on the MFT25 dataset. † indicates the integration of Re-ID module. TransF denotes the Transformer-based model. The best two results are bolded and underlined respectively. Same as follows.

Method	IBN	HOTA↑	IDF1↑	MOTA↑	AssA↑	IDs↓
SBS-R50		30.950	39.937	68.780	19.599	713
SBS-R50	✓	30.560	37.919	68.909	19.010	659
SBS-R101		30.270	38.104	68.796	18.599	678
SBS-R101	✓	30.684	39.996	68.912	19.134	638
SBS-S50		32.705	41.727	69.167	21.701	562
SBS-S50	✓	32.412	40.977	69.030	21.183	558
SBS-S101		**33.842**	**43.748**	69.043	**23.154**	**550**
SBS-S101	✓	31.900	40.201	**69.212**	20.610	584

Table 3: Results on different Re-ID models with standard Kalman Filter and IoU association.

stability, and general applicability across challenging low-light underwater sequences.

Method	HOTA↑	IDF1↑	MOTA↑	AssA↑	IDs↓	IDFP↓
Center	28.865	37.348	66.313	17.410	1273	88585
IoU	32.790	40.098	68.839	21.573	579	84648
CIoU	30.720	39.598	67.425	19.325	727	87422
DIoU	30.764	39.575	67.519	19.326	728	87344
HMIoU	32.258	38.421	68.905	20.936	613	85924
GIoU	32.885	39.957	68.798	21.686	573	84896
FishIoU	33.351	41.717	68.450	22.425	607	83111
FishIoU†	**33.581**	**43.268**	**68.989**	**22.779**	**547**	**78473**

Table 4: Ablation study comparing different IoU for association cost calculation. Center and IoU represent the center points distance and the standard IoU, respectively.

Table 4 presents a comparison of association metrics. Our FishIoU achieves the best results, with 33.4 HOTA and 41.7 IDF1, outperforming standard IoU variants (Zheng et al. 2021, 2020; Yang et al. 2024; Rezatofighi et al. 2019). Incorporating the Re-ID module further boosts performance across all metrics, highlighting FishIoU's effectiveness in handling fish-specific morphology and motion. Table 5 evaluates different motion models. The UKF consistently surpasses the standard Kalman Filter (KF), Adaptive Kalman Filter (AKF), and Strong Tracking Filter (STF) under both HMIoU (Yang et al. 2024) and FishIoU association, confirming the advantage of non-linear modeling for fish tracking. The best performance is achieved by UKF with FishIoU and Re-ID, reaching 34.1 HOTA and 44.6 IDF1.

Method	HMI	FiI	HOTA↑	IDF1↑	MOTA↑	AssA↑	IDs↓
KF	✓		32.258	38.421	68.905	20.936	613
AKF	✓		28.769	33.689	67.827	16.954	1031
STF	✓		31.105	36.911	**69.161**	19.445	667
UKF	✓		32.406	38.408	68.933	21.096	609
UKF†	✓		**33.737**	**43.880**	69.057	**23.063**	**528**
KF		✓	33.051	41.041	68.503	22.022	612
AKF		✓	22.551	24.017	65.535	10.682	2368
STF		✓	31.601	38.137	68.694	20.153	663
UKF		✓	33.201	41.644	68.451	22.261	609
UKF†		✓	**34.067**	**44.643**	**68.958**	**23.594**	**544**

Table 5: Ablation study on different motion models and association IoUs. HMI and FiI denote the use of HMIoU and FishIoU, respectively.

Conclusion

In this paper, we introduce a unified underwater MFT benchmark and a specialized tracking framework for fish morphology and erratic swimming patterns. Our baseline achieves state-of-the-art performance with 34.1 HOTA, significantly outperforming other trackers. Statistical analysis reveals fundamental differences between fish and terrestrial tracking scenarios, highlighting the necessity for specialized underwater approaches. However, significant challenges remain in handling visually similar fish appearances, extreme density scenarios, and highly erratic swimming patterns.

Acknowledgments

The paper is supported in part by Beijing Smart Agriculture Innovation Consortium Project (BAIC10-2025). The authors gratefully acknowledge National Innovation Center for Digital Fishery - China Agricultural University, Key Laboratory of Agricultural Informatization Standardization - MARA, P. R. China, Key Laboratory of Smart Farming Technologies for Aquatic Animals and Livestock - MARA, P. R. China, National Innovation Center for Digital Agricultural Products Circulation - MARA, P. R. China, and State Key Laboratory of Efficient Utilization of Agricultural Water Resources - China Agricultural University. Weiran Li gratefully acknowledges financial support from the China Scholarship Council (No. 202406350102).

References

Aharon, N.; Orfaig, R.; and Bobrovsky, B.-Z. 2022. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*.

Anjum, S.; and Gurari, D. 2020. CTMC: Cell tracking with mitosis detection dataset challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 982–983.

Bernardin, K.; and Stiefelhagen, R. 2008. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008: 1–10.

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, 3464–3468. IEEE.

Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; and Kitani, K. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9686–9696.

Cao, X.; Zheng, Y.; Yao, Y.; Qin, H.; Cao, X.; and Guo, S. 2025. TOPIC: A Parallel Association Paradigm for Multi-Object Tracking under Complex Motions and Diverse Scenes. *IEEE Transactions on Image Processing*.

Cui, M.; Liu, X.; Liu, H.; Zhao, J.; Li, D.; and Wang, W. 2024. Fish Tracking, Counting, and Behaviour Analysis in Digital Aquaculture: A Comprehensive Review. *arXiv preprint arXiv:2406.17800*.

Dawkins, M.; Prior, J.; Lewis, B.; Faillettaz, R.; Banez, T.; Salvi, M.; Rollo, A.; Simon, J.; Campbell, M.; Lucero, M.; et al. 2024. FishTrack23: An Ensemble Underwater Dataset for Multi-Object Tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7167–7176.

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

Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; and Leal-Taixé, L. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fischer, T.; Huang, T. E.; Pang, J.; Qiu, L.; Chen, H.; Darrell, T.; and Yu, F. 2023. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Hassan, S.; Mujtaba, G.; Rajput, A.; and Fatima, N. 2024. Multi-object tracking: a systematic literature review. *Multimedia Tools and Applications*, 83(14): 43439–43492.

He, L.; Liao, X.; Liu, W.; Liu, X.; Cheng, P.; and Mei, T. 2023. Fastreid: A pytorch toolbox for general instance reidentification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9664–9667.

Huang, T.-W.; Hwang, J.-N.; Romain, S.; and Wallace, F. 2018. Fish tracking and segmentation from stereo videos on the wild sea surface for electronic monitoring of rail fishing. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10): 3146–3158.

Jager, J.; Wolff, V.; Fricke-Neuderth, K.; Mothes, O.; and Denzler, J. 2017. Visual fish tracking: Combining a two-stage graph approach with CNN-features. In *OCEANS 2017-Aberdeen*, 1–6. IEEE.

Kay, J.; Kulits, P.; Stathatos, S.; Deng, S.; Young, E.; Beery, S.; Van Horn, G.; and Perona, P. 2022. The caltech fish counting dataset: A benchmark for multiple-object tracking and counting. In *European Conference on Computer Vision*, 290–311. Springer.

Lei, Y.; Zhu, H.; Yuan, J.; Xiang, G.; Zhong, X.; and He, S. 2024. DenseTrack: Drone-based Crowd Tracking via Density-aware Motion-appearance Synergy. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2050–2058.

Li, G.; Jian, Y.; Jian, Y.; Yan, Y.; Yan, Y.; Wang, H.; and Wang, H. 2024a. GLATrack: Global and Local Awareness for Open-Vocabulary Multiple Object Tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2457–2466.

Li, W.; Li, F.; and Li, Z. 2022. CMFTNet: Multiple fish tracking based on counterpoised JointNet. *Computers and electronics in agriculture*, 198: 107018.

Li, W.; Liu, Y.; Wang, W.; Li, Z.; and Yue, J. 2024b. TFMFT: Transformer-based multiple fish tracking. *Computers and Electronics in Agriculture*, 217: 108600.

Li, X.; Wei, Z.; Huang, L.; Nie, J.; Zhang, W.; and Wang, L. 2018. Real-time underwater fish tracking based on adaptive multi-appearance model. In *2018 25th IEEE international conference on image processing (ICIP)*, 2710–2714. IEEE.

Liu, C.; Li, H.; and Wang, Z. 2023. FastTrack: A Highly Efficient and Generic GPU-Based Multi-object Tracking Method with Parallel Kalman Filter. *International Journal of Computer Vision*, 1–21.

Liu, Y.; Wu, J.; and Fu, Y. 2023. Collaborative Tracking Learning for Frame-Rate-Insensitive Multi-Object Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9964–9973.

Liu, Z.; Wang, X.; Wang, C.; Liu, W.; and Bai, X. 2025. Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *IEEE Transactions on Circuits and Systems for Video Technology*.

Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; and Leibe, B. 2021. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129: 548–578.

Luo, R.; Song, Z.; Ma, L.; Wei, J.; Yang, W.; and Yang, M. 2024. Diffusiontrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3991–3999.

Maggiolino, G.; Ahmad, A.; Cao, J.; and Kitani, K. 2023. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International conference on image processing (ICIP)*, 3025–3029. IEEE.

Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; and Feichtenhofer, C. 2022. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8844–8854.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, 464–479.

Pedersen, M.; Haurum, J. B.; Bengtson, S. H.; and Moeslund, T. B. 2020. 3d-zef: A 3d zebrafish tracking benchmark dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2426–2436.

Pedersen, M.; Lehotskỳ, D.; Nikolov, I.; and Moeslund, T. B. 2023. Brackishmot: The brackish multi-object tracking dataset. In *Scandinavian Conference on Image Analysis*, 17–33. Springer.

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

Seidenschwarz, J.; Brasó, G.; Serrano, V. C.; Elezi, I.; and Leal-Taixé, L. 2023. Simple cues lead to a strong multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13813–13823.

Shevchenko, V.; Eerola, T.; and Kaarna, A. 2018. Fish detection from low visibility underwater videos. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 1971–1976. IEEE.

Sun, P.; Cao, J.; Jiang, Y.; Yuan, Z.; Bai, S.; Kitani, K.; and Luo, P. 2022. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20993–21002.

Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; and Luo, P. 2020. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*.

Wang, H.; Zhang, S.; Zhao, S.; Wang, Q.; Li, D.; and Zhao, R. 2022. Real-time detection and tracking of fish abnormal behavior based on improved YOLOV5 and SiamRPN++. *Computers and Electronics in Agriculture*, 192: 106512.

Xiao, C.; Cao, Q.; Luo, Z.; and Lan, L. 2024. Mambatrack: a simple baseline for multiple object tracking with state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4082–4091.

Xu, Y.; Ban, Y.; Delorme, G.; Gan, C.; Rus, D.; and Alameda-Pineda, X. 2022. TransCenter: Transformers with dense representations for multiple-object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 45(6): 7820–7835.

Yang, M.; Han, G.; Yan, B.; Zhang, W.; Qi, J.; Lu, H.; and Wang, D. 2024. Hybrid-sort: Weak cues matter for online multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6504–6512.

Yi, K.; Luo, K.; Luo, X.; Huang, J.; Wu, H.; Hu, R.; and Hao, W. 2024. Ucmctrack: Multi-object tracking with uniform camera motion compensation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 6702–6710.

Zeng, Y.; Yang, X.; Pan, L.; Zhu, W.; Wang, D.; Zhao, Z.; Liu, J.; Sun, C.; and Zhou, C. 2023. Fish school feeding behavior quantification using acoustic signal and improved Swin Transformer. *Computers and Electronics in Agriculture*, 204: 107580.

Zhang, C.; Liu, L.; Huang, G.; Wen, H.; Zhou, X.; and Wang, Y. 2024. Webuot-1m: Advancing deep underwater object tracking with a million-scale benchmark. *arXiv preprint arXiv:2405.19818*.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, 1–21. Springer.

Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129: 3069–3087.

Zhao, Z.; Wu, Z.; Zhuang, Y.; Li, B.; and Jia, J. 2022. Tracking objects as pixel-wise distributions. In *European Conference on Computer Vision*, 76–94. Springer.

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

Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; and Zuo, W. 2021. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE transactions on cybernetics*, 52(8): 8574–8586.