

Article

Depth-Prior-Guided Detection and Turning-Aware Association for Multi-Object Tracking of Marine Fish in Aquaculture Tanks

Zhao Li ¹, Feifan Yan ¹ , Guangpeng Huang ¹, Longhua Qiu ¹, Yunrong Yan ^{2,3} and Siwei Zhou ^{1,*}

¹ College of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, China; zhaoli@gdou.edu.cn (Z.L.); yanff927@stu.gdou.edu.cn (F.Y.)

² College of Fisheries, Guangdong Ocean University, Zhanjiang 524088, China; yanyr@gdou.edu.cn

³ Guangdong Provincial Engineering and Technology Research Center of Far Sea Fisheries Management and Fishing of South China Sea, Guangdong Ocean University, Zhanjiang 524088, China

* Correspondence: siweizhou@gdou.edu.cn

Abstract

Continuous tracking of individual fish provides a basis for movement analysis and visual monitoring in marine aquaculture. Underwater image degradation, similar appearance, and non-rigid motion complicate fish detection and identity preservation. This study develops a tracking framework for golden pompano (*Trachinotus ovatus*) and large yellow croaker (*Larimichthys crocea*) recorded in a land-based aquaculture tank. The framework combines a depth-structure-prior-enhanced RT-DETRv4 detector with the Scale-aware and Unscented Tracker (SU-T). A frozen Depth Anything 3 (DA3) model provides supervision based on physics-inspired structural cues derived from relative depth during training, while a lightweight student head predicts the prior without online teacher inference. Fish Turning–Deformation Coupling Consistency (FTDC) uses the empirical relationship between short-term motion-direction change and observable bounding-box variation in trajectory association. On the self-collected dataset, the final detector increased average precision over intersection-over-union thresholds of 0.50–0.95 (AP50:95) from 0.837 to 0.851. With identical detection inputs, SU-T+FTDC achieved 88.3% multiple-object-tracking accuracy, 77.4% identification F1 score, and 62.8% higher-order tracking accuracy; identity switches decreased from 66 to 60 and trajectory fragmentations from 86 to 78 relative to SU-T. These results support improved fish localization and identity continuity, providing a trajectory basis for subsequent movement and behavior analyses in aquaculture.

Keywords: marine aquaculture fish; multi-object tracking; depth-structure prior; feature enhancement; non-rigid motion; data association

Key Contribution: This study combines lightweight depth-prior-guided fish detection with a turning-conditioned geometric constraint for trajectory association, improving localization and identity continuity in marine aquaculture videos.



Academic Editor: Marina Paolucci

Received: 6 September 2026

Revised: 5 October 2026

Accepted: 6 October 2026

Published: 9 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Continuous observation of cultured fish provides information relevant to movement analysis, behavioral monitoring, and aquaculture management [1,2]. Video-based multi-object tracking links successive observations of the same fish and provides individual trajectories for these applications. In marine aquaculture, however, water scattering, illumination attenuation, suspended particles, and background disturbances reduce fish–background contrast and can cause missed detections and unstable localization [3]. Similar appearance

among neighboring fish further complicates identity preservation when individuals cross or overlap. Turning, tail beating, and body bending can also change a fish's projected shape without a change in identity. FishTrack23 and MFT25 document these challenges, including dynamic backgrounds, scale variation, overlap, and nonlinear motion [4,5]. Reliable fish monitoring therefore requires both accurate frame-level localization and consistent identity association across successive observations.

To obtain reliable frame-level fish observations, underwater object detection studies have investigated image enhancement, multi-scale feature fusion, attention mechanisms, and detection-network optimization [6,7]. MarineYOLO and AquaYOLO address small targets, low contrast, and background interference in underwater environments [8,9]. A recent YOLO26n fish detector introduces a lightweight convolutional attention module to strengthen feature discrimination under difficult underwater imaging conditions [10]. The RT-DETR family of end-to-end detectors balances detection accuracy and inference efficiency and can provide frame-wise fish candidate boxes for tracking-by-detection frameworks [11]. These methods mainly learn texture and semantic features from red–green–blue (RGB) images; however, water scattering and illumination attenuation cause such appearance cues to vary with spatial position and imaging distance. Monocular relative-depth responses describe the spatial arrangement of targets and background. Monocular depth foundation models such as Depth Anything 3 (DA3) exhibit strong scene-structure representation capability [12], but their outputs are not specifically designed for underwater fish detection, and directly integrating them online into a detector introduces additional computational overhead. These observations motivate using depth-derived cues to guide detection-feature encoding under underwater appearance degradation.

To maintain individual identities across frames, tracking-by-detection methods typically associate trajectories with candidate detections using motion prediction, bounding-box overlap, detection confidence, or appearance similarity [13,14]. SORT and DeepSORT perform online association using Kalman filtering and deep appearance features, respectively [15,16]. ByteTrack improves the utilization of low-confidence detections by jointly associating high- and low-confidence bounding boxes [17]. OC-SORT and BoT-SORT enhance state estimation and association robustness through observation-centric motion modeling and through motion, appearance, and camera-motion compensation, respectively [18,19].

For fish tracking, the robust fish-school tracking algorithm (FSTA) combines detection correction and underwater-specific association with a re-identification (Re-ID) branch [20], while an improved DeepSORT method strengthens appearance discrimination through deformable convolution, attention, and metric learning [21]. UMOTMA incorporates temporal information through memory aggregation [22]. Other fish trackers employ detector enhancement, nonlinear motion, spatiotemporal fusion, and Transformer-based sequence models [23–28].

Multimodal tracking studies exploit complementary information through modality-aware representations, frequency- and memory-aware modeling, and spatial–temporal interaction [29–31]. The present study explores complementary cues within an RGB-based pipeline, using relative-depth supervision for detection and motion–geometry constraints for association.

At the motion and geometry level, the Scale-aware and Unscented Tracker (SU-T) in MFT25 uses an unscented Kalman filter for nonlinear fish motion and FishIoU for scale and shape variation [5]. Forward-orientation estimation has also been explored in angle-aware fish tracking [32]. These studies motivate complementing motion and shape cues with a geometric tolerance conditioned on short-term trajectory-direction change. During turning or body bending, area and aspect ratio may change, while fish identity remains the same.

A fixed geometric penalty may reject valid candidates, while relaxing geometric constraints may increase incorrect matches after fish crossing.

For continuous fish monitoring, these observations motivate two complementary improvements: more reliable detection under degraded underwater appearance and an association that accommodates the geometric changes accompanying fish motion. This study addresses both needs using videos of golden pompano (*Trachinotus ovatus*) and large yellow croaker (*Larimichthys crocea*) collected in a land-based aquaculture tank. DA3-derived relative-depth structure guides multi-scale feature gating and attention in RT-DETRv4. A lightweight student head learns the task-related structural responses so that the final detector operates without online DA3 inference. For association, Fish Turning–Deformation Coupling Consistency (FTDC) conditions an allowable envelope of bounding-box area and aspect ratio changes on short-term turning amplitude and adds a soft penalty for candidates outside this envelope to SU-T matching. The evaluation examines fish detection and trajectory identity continuity under the recorded conditions.

The main contributions of this study are as follows:

- (1) A teacher-guided depth-structure-prior detector transfers DA3-derived structural cues to TinyDepthPriorHead. The predicted prior guides RT-DETRv4 through multi-scale residual gating and attention bias, retaining relative-depth guidance in the final detector without online DA3 inference.
- (2) FTDC extends SU-T association with a turning-conditioned envelope for bounding-box area and aspect ratio changes. Candidate matches are softly penalized for geometric variation exceeding this envelope, complementing SU-T's existing nonlinear motion prediction and FishIoU geometry matching.

2. Materials and Methods

2.1. Data Acquisition, Annotation, and Dataset Partitioning

Data were collected at the Waipo Base of Guangdong Hetai Fishery Technology Co., Ltd., Yangjiang, China. The experimental scene consisted of a land-based aquaculture tank measuring 4 m × 4 m with a water depth of approximately 1 m. Videos were recorded using an underwater camera (model IPC5MPW-PBX5; Shenzhen Shuimitan Technology Co., Ltd., Shenzhen, China) at 1920 × 1080 pixels and 30 frames per second (fps). The camera was mounted at the midpoint of one sidewall, with its lens center approximately 0.15 m above the tank bottom, and was directed toward the geometric center of the tank, approximately 2 m away in the horizontal plane. Recordings were made under natural daylight without supplementary artificial illumination. To improve the model's adaptability to different fish appearances, scales, and motion states, two batches of cultured fish were selected. The first batch consisted of 30 golden pompano (*Trachinotus ovatus*; GP), with an average body length of approximately 15 cm; the second batch consisted of 30 large yellow croaker (*Larimichthys crocea*; LC), with an average body length of approximately 11 cm.

To reduce repetitive annotation while retaining the temporal information required for tracking, each selected 30 s clip was uniformly subsampled at a fixed interval before annotation; frames were not selected manually. An effective rate of 5, 10, or 15 fps was assigned to each sequence according to fish movement speed and the observed redundancy between adjacent frames. These rates corresponded to retaining every sixth, third, or second frame, respectively, from the original 30 fps recording. Faster-motion or less-redundant sequences were retained at a higher rate, whereas slower-motion or more-redundant sequences were retained at a lower rate. The resulting corpus contained 10 sequences at 5 fps, four at 10 fps, and seven at 15 fps. Related underwater fish video datasets and behavior analysis studies have likewise reduced annotation or processing rates to balance labeling effort, computational cost, temporal coverage, and preservation of fish

motion [4,33]. Fish-tracking experiments at 5, 10, and 20 fps have further shown that tracking performance and tracker settings can depend on the effective frame rate [34]. Accordingly, the original acquisition rate and effective post-sampling rate are reported separately in this study.

The sampled clips were organized as a multi-object-tracking dataset in MOTChallenge format. The GP subset contained 16 sequences and 3600 annotated frames, whereas the LC subset contained five sequences and 2250 annotated frames; the 21 clips therefore provided a total annotated duration of 10.5 min. Fish targets were manually annotated using DarkLabel (v2.4) with rectangular bounding boxes, and the same fish was assigned a consistent target identity within each sequence. The annotations were first stored as DarkLabel-style CSV files and then converted uniformly to MOTChallenge format. A Common Objects in Context (COCO)-format detection dataset was generated from the same sampled frames, treating each frame as one detection sample and all fish as a single target class. The detection dataset retained the same sequence boundaries and subset assignments as the MOT dataset.

The source material comprised two continuous original videos recorded on the same day: an approximately 2 h morning video and a separate approximately 30 min afternoon video. Non-overlapping 30 s clips were randomly selected from these videos. All training and validation clips came from the morning video, whereas all test clips came from the afternoon video. Training and validation were assigned at the clip level within their shared source video; the test set was held out at the original-video level. The training set contained 14 sequences (11 GP and three LC sequences), comprising 3750 images and 20,250 bounding-box annotations. The validation set contained three sequences (two GP and one LC sequence), comprising 750 images and 3767 annotations. The test set contained four sequences (three GP and one LC sequence), comprising 1350 images and 8996 annotations; both species and all three effective frame rates were represented in the test set. In total, the dataset consisted of 21 sequences, 5850 annotated images, and 33,013 bounding boxes. Bounding-box counts denote frame-level target instances rather than the number of independent fish.

2.2. Overall Framework

This study improves marine aquaculture fish multi-object tracking from two aspects: the detection front end and trajectory association. The overall framework is shown in Figure 1. First, the depth-structure-prior-enhanced detector extracts fish candidate boxes from the current video frame. The detection results are then input to the SU-T tracker, where motion-state prediction, the baseline association cost, and the proposed FTDC turning-deformation coupling cost jointly perform trajectory-detection matching. Finally, the track management module outputs fish-tracking results with continuous identity labels.

Given the input image at frame t , it can be expressed as follows:

$$I_t \in \mathbb{R}^{H \times W \times 3} \quad (1)$$

where H and W denote the image height and width, respectively. The detector generates the fish candidate set for the current frame as follows:

$$\mathcal{D}_t = \{d_i^t = (b_i^t, c_i^t, q_i^t)\}_{i=1}^{N_t} \quad (2)$$

where the three components of each candidate detection in Equation (2) denote the bounding box, class label, and detection confidence, respectively. N_t is the number of candidate detections in frame t . All fish targets are treated as a single class; therefore, subsequent track-

ing primarily relies on bounding-box geometric states, motion information, and association costs for identity matching.

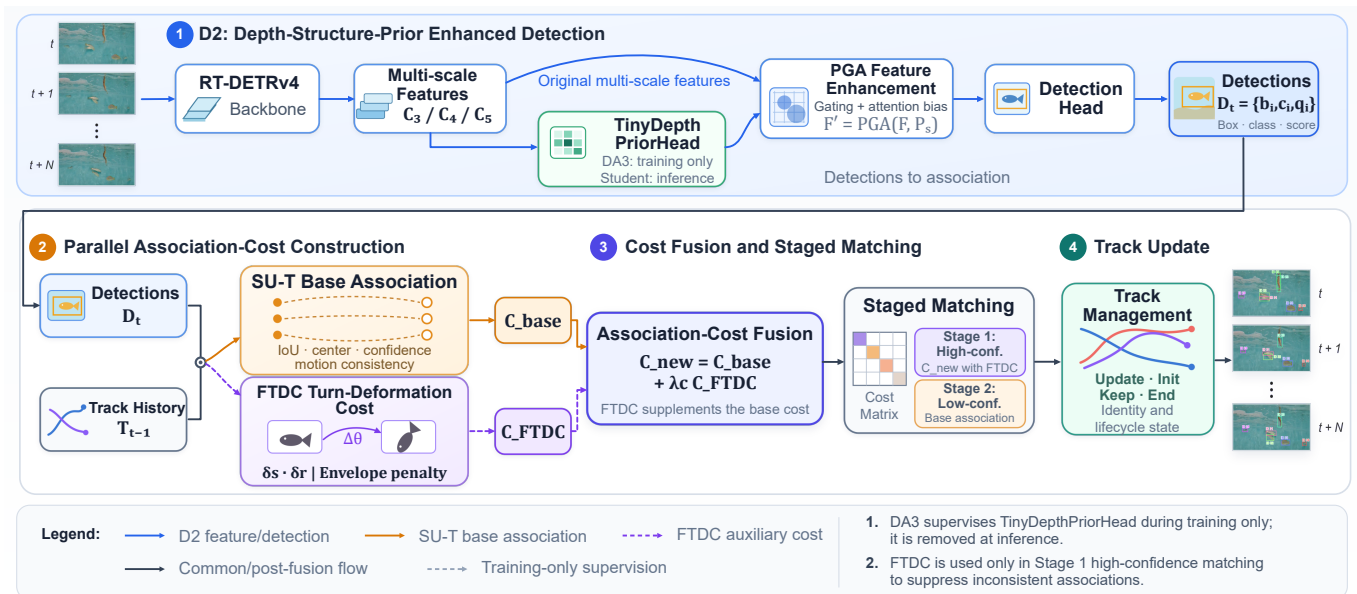


Figure 1. Overall framework of the proposed multi-object-tracking method for marine aquaculture fish. Abbreviations: DA3, Depth Anything 3; PGA, Prior-Guided Attention; FTDC, Fish Turning–Deformation Coupling Consistency; SU-T, Scale-aware and Unscented Tracker. Arrow colors and line styles are defined in the figure legend.

The officially pretrained RT-DETRv4-HGNetv2-M is used as the baseline detector [35], and the original vision foundation model (VFM) teacher branch and its distillation loss are not enabled during downstream training. To introduce underwater structural information, three detection configurations, D0, D1, and D2, are defined. D0 is the baseline without an additional depth-structure prior; D1 uses a frozen DA3 online to generate structural priors; and D2 uses the frozen DA3 during training to supervise TinyDepthPriorHead in learning task-related structural responses, following a teacher–student knowledge-transfer strategy [36,37] while retaining only the student-prior branch during inference. D1 is used to analyze the effect of the online prior, whereas D2 is the final inference configuration. Details are provided in Section 2.3.

After the detector outputs candidate boxes, each bounding box is converted into a geometric observation containing center position, area, and aspect ratio:

$$z_i^t = [c_{x,i}^t, c_{y,i}^t, s_i^t, r_i^t]^\top \quad (3)$$

where the four components in Equation (3) denote the horizontal coordinate of the bounding-box center, vertical coordinate of the center, bounding-box area, and bounding-box aspect ratio, respectively. The center position is used for motion-state updating and short-term direction estimation, while the area and aspect ratio record observable bounding-box changes that may accompany turning or posture variation but can also reflect occlusion, localization error, or motion toward or away from the camera.

During tracking, SU-T first predicts the target position in the current frame based on the historical trajectory state and constructs the baseline association cost matrix using motion information, spatial overlap, and detection confidence. For the j -th trajectory and the i -th candidate detection, this study further estimates the short-term turning amplitude from the historical trajectory direction and the current candidate displacement, and calculates

the FTDC coupling cost using changes in area and aspect ratio. The final association cost is expressed as follows:

$$C_{\text{new}} = C_{\text{base}} + \lambda_c C_{\text{FTDC}} \quad (4)$$

where the scalar coefficient in Equation (4) denotes the weight of the FTDC coupling cost. FTDC does not replace the original motion-prediction or spatial-matching mechanisms of SU-T; instead, it imposes an additional soft penalty during candidate matching on associations that are inconsistent with the fish turning–deformation coupling relationship.

Based on the new association cost matrix, the tracker matches trajectories with detections and performs management operations including trajectory initialization, state updating, lost-track retention, and termination. The final tracking result at frame t can be expressed as follows:

$$\mathcal{T}_t = \left\{ \left(\hat{b}_j^t, id_j, \hat{p}_j^t \right) \right\}_{j=1}^{M_t} \quad (5)$$

where each triplet in Equation (5) contains the bounding box of a valid trajectory in the current frame, its trajectory identity, and the confidence of the corresponding detection, respectively. M_t is the number of valid trajectories in frame t .

Accordingly, the final inference pipeline can be summarized as D2 detection, SU-T baseline association, FTDC coupling cost correction, and trajectory output; D1 is used only as a reference configuration for the detector with an online DA3 prior.

2.3. Depth-Structure-Prior-Enhanced Detector

The officially pretrained RT-DETRv4-HGNetv2-M is adopted as the baseline detection network [35]. The network consists of a backbone, a Hybrid Encoder, and a detection decoder. A DA3-based depth-structure-prior enhancement mechanism is introduced at the multi-scale feature-encoding stage without changing the original detection-decoding process. The backbone extracts multi-scale visual features from the input image as follows:

$$\{C_3, C_4, C_5\} = \mathcal{B}(I) \quad (6)$$

The Hybrid Encoder further performs cross-scale interaction and semantic fusion on the multi-scale features, which can be expressed as follows:

$$Z = \mathcal{E}(C_3, C_4, C_5) \quad (7)$$

The detection decoder generates object detection results from the encoder output as follows:

$$Y = \mathcal{D}_{\text{det}}(Z) \quad (8)$$

where $\mathcal{B}$ denotes the RT-DETRv4 backbone; C_3 , C_4 , and C_5 denote multi-scale features with different spatial resolutions and semantic levels; $\mathcal{E}$ denotes the Hybrid Encoder; Z denotes the encoder output; and $\mathcal{D}_{\text{det}}$ denotes the detection decoder. For marine aquaculture images, water scattering, illumination attenuation, and low contrast weaken local texture and boundary responses derived solely from RGB appearance information. Therefore, rather than changing the original RT-DETRv4 decoding process, this study introduces additional priors constructed from relative depth structure at the multi-scale feature-encoding stage, allowing the detector to jointly use RGB semantic features and depth-derived cues describing relative scene structure.

D1 and D2 share the same prior-guided feature-enhancement mechanism and differ only in how the structural prior is generated: D1 directly uses the online frozen DA3 prior, whereas D2 uses the student prior predicted by TinyDepthPriorHead. In this study, the

multi-scale prior gating and attention-bias injection mechanisms are collectively referred to as Prior-Guided Attention (PGA).

2.3.1. D1: Online DA3 Depth-Structure-Prior Enhancement

D1 evaluates the contribution of DA3 relative-depth structure to fish detection. During training and inference, the frozen DA3 processes each RGB frame to provide relative spatial information between targets and background. Its parameters are excluded from backpropagation. D1 retains the original RT-DETRv4 detection loss without an additional depth-prior alignment loss. Given an input image, DA3 outputs the relative-depth response as follows:

$$D = \mathcal{T}_{DA3}(I) \quad (9)$$

The depth response is min–max-normalized independently for each image to [0, 1]:

$$\hat{D}(x) = \frac{D(x) - D_{\min}}{D_{\max} - D_{\min} + \varepsilon} \quad (10)$$

where $D_{\max}$ and $D_{\min}$ denote the maximum and minimum values, respectively, of the DA3 depth response D in the current image, and ε is a small constant used to prevent division by zero. The normalized response $\hat{D}(x)$ is used to construct task-related structural priors.

During underwater imaging, target propagation distance is correlated with phenomena such as direct-light attenuation and background scattering. More rigorous underwater image formation models indicate that direct-signal attenuation and backscattering are governed by different distance-dependent parameters [38,39]. These physical relationships motivate the shapes of the relative-depth mappings used below. A simplified underwater image formation relationship can be expressed as follows:

$$I(x) = J(x)T(x) + B^{\infty}(1 - T(x)) \quad (11)$$

where $I(x)$ denotes the observed image intensity, $J(x)$ denotes the ideal undegraded image radiance, B^{∞} denotes the ambient-light or background-scattering term, and $T(x)$ denotes the medium transmittance. The corresponding distance-dependent transmission relationship can be expressed as follows:

$$T(x) = \exp(-\beta d(x)) \quad (12)$$

In the illustrative physical model, β denotes the attenuation coefficient and $d(x)$ denotes scene depth. An attenuation-shaped structural cue is defined from normalized DA3 depth $\hat{D}(x)$ (Figure 2) as follows:

$$T_d(x) = \exp(-\beta_t \hat{D}(x)) \quad (13)$$

A complementary depth-dependent structural cue is defined as follows:

$$S_b(x) = 1 - \exp(-\beta_s \hat{D}(x)) \quad (14)$$

A composite structural cue is defined as follows:

$$R(x) = T_d(x) \odot (1 - S_b(x)) \quad (15)$$

Together, these three responses constitute the physics-inspired structural-prior set:

$$P = \{T_d, S_b, R\} \quad (16)$$

Here, β_t and β_s are dimensionless scale factors controlling the exponential mappings, and $\odot$ denotes element-wise multiplication. The three cues are derived from the same normalized relative-depth map, with $R(x) = \exp[-(\beta_t + \beta_s)\hat{D}(x)]$. They form a heuristic structural prior for subsequent feature gating and attention-bias generation.

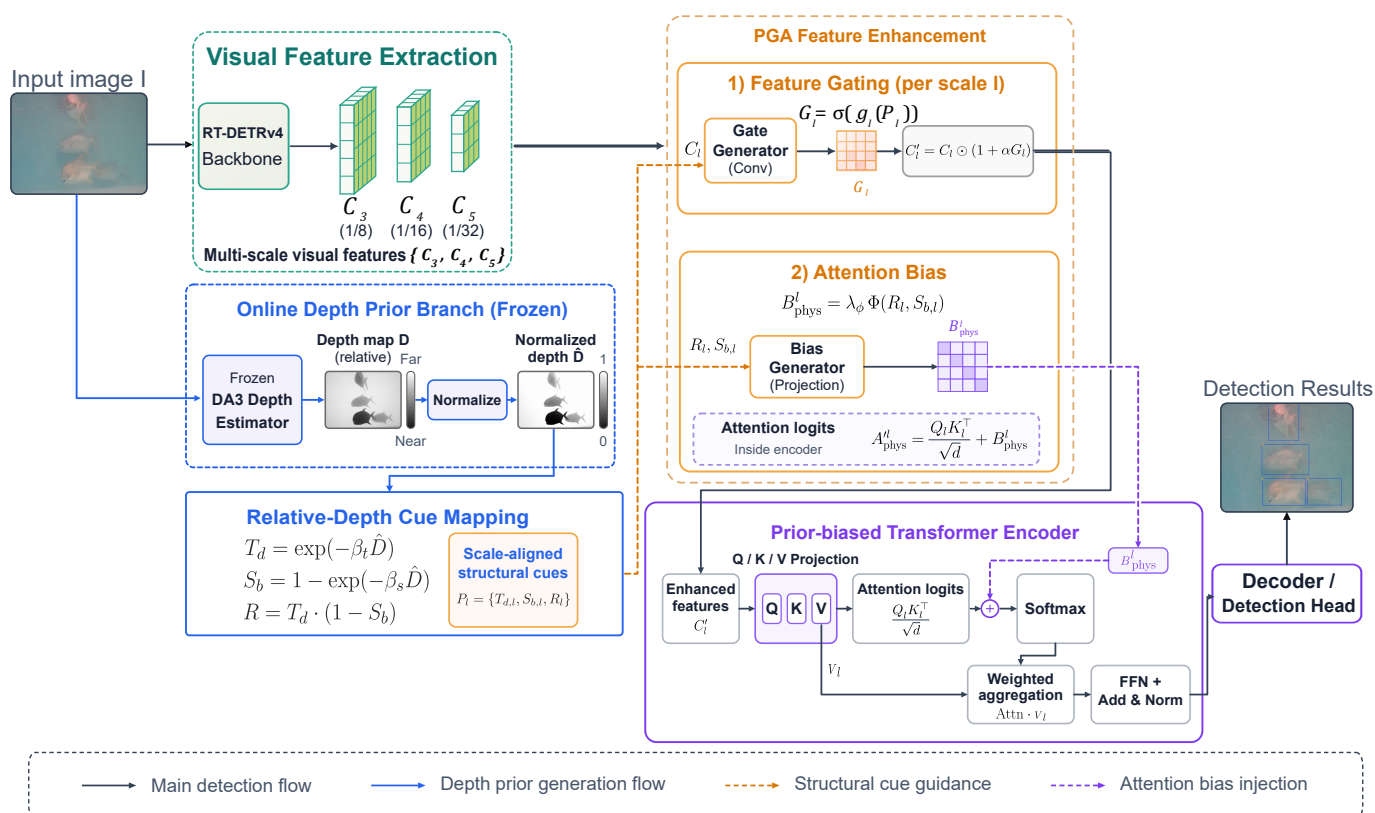


Figure 2. Online DA3 depth-structure prior generation and feature enhancement through multi-scale gating and attention-bias injection. Green, blue, orange, and purple outlines group visual feature extraction, depth-prior generation, prior-guided feature enhancement, and transformer encoding, respectively. Dashed outlines delimit grouped branches or modules.

The RT-DETRv4 encoder receives multi-scale visual features, whereas the physics-inspired structural prior derived from DA3 normalized depth has the same resolution as the input image. Therefore, the structural-prior maps must be aligned with different feature-map scales. For the feature C_l at the l -th level, whose spatial size is $H_l \times W_l$, the corresponding scale-aligned prior can be expressed as follows:

$$P_l = \text{Resize}(P, H_l, W_l) = \{T_{d,l}, S_{b,l}, R_l\} \quad (17)$$

where P_l denotes the structural prior aligned with the feature at the l -th level. A convolutional gating module is then used to generate the gating weights:

$$G_l = \sigma(g_l(P_l)) \quad (18)$$

where $g_l()$ denotes the gating-generation function and σ denotes the Sigmoid function. Consistent with Figure 2, the gating weight G_l is generated from the scale-aligned structural prior P_l . To avoid excessive suppression of feature magnitudes by direct multiplication, residual feature modulation is adopted:

$$C'_l = C_l \odot (1 + \alpha G_l) \quad (19)$$

where α denotes the gating-strength coefficient and C'_l denotes the multi-scale feature enhanced by the structural prior. This design retains the original RGB semantic features while moderately enhancing or suppressing different regions according to the task-related structural prior.

In addition to multi-scale feature gating, the structural prior is further introduced into the Transformer attention computation. Residual gating mainly adjusts the response magnitude of features in different spatial regions, whereas the attention bias affects spatial-relation modeling within the Transformer. The standard attention logits can be expressed as follows:

$$A_l = \frac{Q_l K_l^\top}{\sqrt{d}} \quad (20)$$

where Q_l and K_l denote the query and key, respectively, corresponding to the feature at the l -th level, and d denotes the channel dimension. The composite cue R_l and complementary cue $S_{b,l}$, both derived from relative depth, are mapped to an attention bias:

$$B_{\text{phys}}^l = \lambda_\phi \Phi(R_l, S_{b,l}) \quad (21)$$

where $\Phi()$ denotes the mapping function that converts the two depth-derived cues into an attention-bias matrix, and λ_ϕ denotes the attention-bias strength. To prevent the bias magnitude from becoming excessively large, the attention bias is first clipped:

$$\bar{B}_{\text{phys}}^l = \text{Clip}(B_{\text{phys}}^l, -b, b) \quad (22)$$

where b denotes the bias-clipping threshold, and the left-hand side of Equation (22) denotes the clipped structural-prior attention bias. After the clipped structural bias is added, the attention logits become:

$$A_{\text{phys}}^l = \frac{Q_l K_l^\top}{\sqrt{d}} + \bar{B}_{\text{phys}}^l \quad (23)$$

The corresponding attention output is:

$$O_{\text{phys}}^l = \text{Softmax}(A_{\text{phys}}^l) V_l \quad (24)$$

where V_l denotes the value. PGA uses depth-structure information only through residual gating and the structural-prior attention bias.

2.3.2. D2: DA3-Guided TinyDepthPriorHead Prior Learning

D1 still requires DA3 to run frame by frame during inference. To avoid this additional computation while retaining its task-related structural information, D2 uses structural priors generated by the frozen DA3 during training to supervise TinyDepthPriorHead, which predicts the student prior from RT-DETRv4 multi-scale features. During inference, DA3 is removed and only TinyDepthPriorHead and PGA are retained. The D2 training framework is shown in Figure 3.

This study initializes the model with the official pretrained weights but does not enable the original VFM teacher branch during downstream training. The newly introduced frozen DA3 is used only to generate a three-channel task-related structural prior as the supervision target for TinyDepthPriorHead.

In D2, the frozen DA3 is used only as a teacher during training. Its depth response D for an input image I is defined in Equation (9).

The teacher physics-inspired structural prior is generated by the Depth Prior Adapter and detached to prevent teacher-branch parameter updates:

$$P_t = \text{Detach}(\mathcal{A}_{\text{phy}}(D)) \quad (25)$$

where $\mathcal{A}_{\text{phy}}(\cdot)$ denotes the non-parametric depth-prior mapping defined by Equations (13)–(15), which has no trainable parameters. This operation ensures that the DA3 teacher branch is excluded from backpropagation, and the teacher prior P_t serves only as a stable structural supervision signal for constraining the student branch.

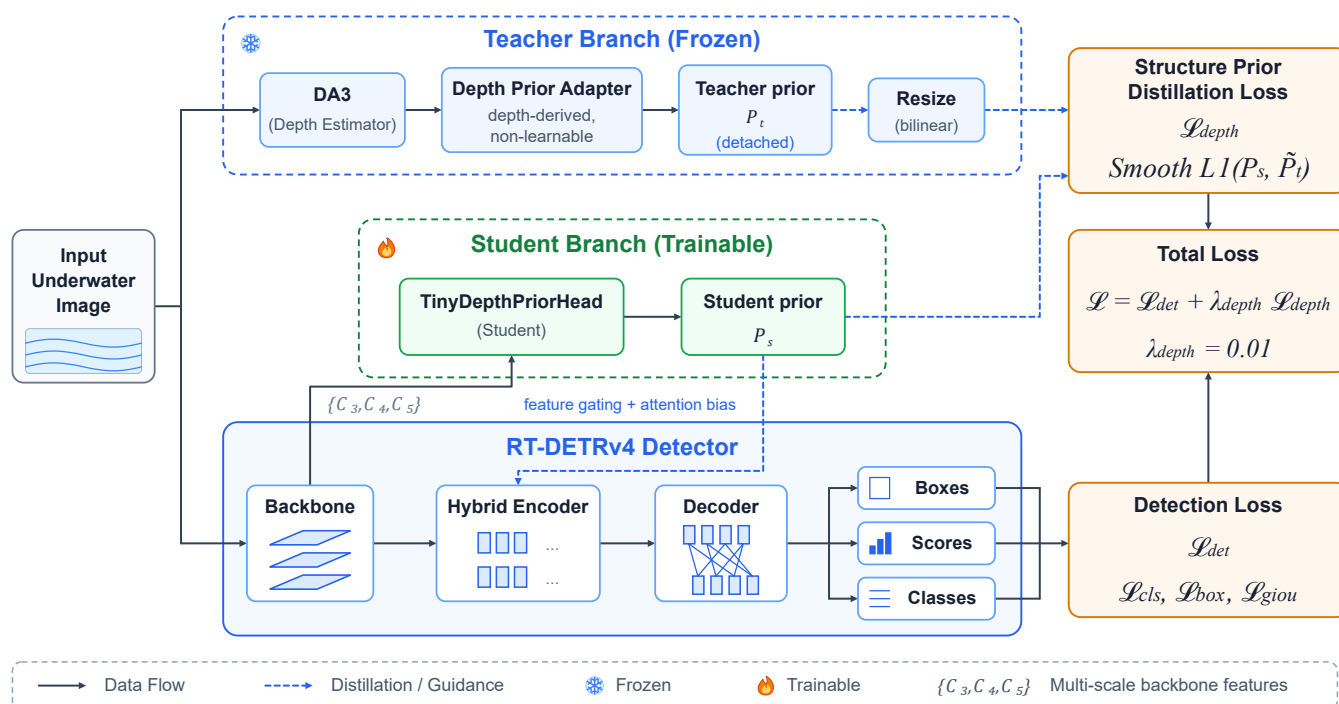


Figure 3. Training framework of DA3-guided TinyDepthPriorHead for task-related structure prior learning. The fixed Depth Prior Adapter generates heuristic supervision targets from DA3 relative depth. Blue and green dashed boxes denote the frozen teacher and trainable student branches, respectively. The blue solid box encloses the detector, and orange boxes denote training losses.

TinyDepthPriorHead predicts the task-related structural prior from the existing multi-scale features of the RT-DETRv4 backbone; therefore, an independent depth foundation model does not need to be run again during inference. The backbone features C_3, C_4, C_5 are defined in Equation (6).

The student branch first performs channel reduction on the features at each scale. To avoid confusion with the enhanced feature C'_l in Figure 2, the reduced-dimensional feature within the student branch is denoted by F_l :

$$F_l = \phi_l(C_l), \quad l \in \{3, 4, 5\} \quad (26)$$

The features at different scales are then resized to a common spatial resolution and fused along the channel dimension:

$$U = \text{Concat}(F_3, \text{Up}(F_4), \text{Up}(F_5)) \quad (27)$$

The fused features are further used to generate the student structural prior:

$$P_s = \sigma(\psi(U)) \quad (28)$$

where $\phi_l()$ denotes 1×1 convolution for channel reduction, $\text{Up}()$ denotes upsampling, $\text{Concat}()$ denotes channel concatenation, and $\psi()$ denotes the fusion function; the output channel number of the final layer of ψ is set to 3. P_s denotes the three-channel task-related depth-prior response learned by TinyDepthPriorHead. Its three channels are aligned with T_d , S_b , and R of the teacher prior P_t .

To enable TinyDepthPriorHead to learn the task-related structural priors provided by the DA3 teacher, a depth-structure-prior distillation loss is introduced. This teacher–student transfer follows the general knowledge distillation paradigm [36,37], but the transferred target here is a task-related structural prior rather than class logits. The teacher prior $\tilde{P}_t$ is first resized to the spatial scale of the student prior P_s :

$$\tilde{P}_t = \text{Resize}(P_t, H_s, W_s) \quad (29)$$

where H_s and W_s denote the spatial dimensions of the student structural prior P_s , and $\tilde{P}_t$ denotes the teacher structural prior after bilinear interpolation. A Smooth L1 loss is then used to constrain the consistency between the student and teacher structural priors:

$$\mathcal{L}_{\text{depth}} = \frac{1}{N} \sum_x \text{SmoothL1}(P_s(x), \tilde{P}_t(x)) \quad (30)$$

where N denotes the number of pixels in the structural-prior map and x denotes the spatial position. The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}} \quad (31)$$

where $\mathcal{L}_{\text{det}}$ denotes the original RT-DETRv4 detection loss, $\mathcal{L}_{\text{depth}}$ denotes the depth-structure-prior distillation loss, and λ_{depth} denotes its weight, which is set to 0.01 in this study.

During training, the teacher and student branches are constructed simultaneously. The teacher branch generates the teacher structural prior P_t using the frozen DA3 and Depth Prior Adapter, whereas the student branch predicts the student structural prior P_s using the RT-DETRv4 backbone and TinyDepthPriorHead and participates in PGA detection-feature enhancement.

During inference, the DA3 teacher model is completely removed, and TinyDepthPriorHead directly generates the student structural prior:

$$P_s = H_{\text{tiny}}(C_3, C_4, C_5) \quad (32)$$

The student structural prior is aligned to multiple scales and then used in PGA feature enhancement:

$$C'_l = G_{\text{enh}}(C_l, P_{s,l}) \quad (33)$$

The final detection output can be expressed as follows:

$$Y = \mathcal{D}_{\text{det}}(\mathcal{E}(\{C'_l\})) \quad (34)$$

where H_{tiny} denotes TinyDepthPriorHead, G_{enh} denotes the structural-prior enhancement module, and $P_{s,l}$ denotes the student structural prior aligned with the feature at the l -th level, $l \in \{3, 4, 5\}$. During D2 inference, DA3 is no longer run; only TinyDepthPriorHead and PGA are retained for detection-feature encoding.

Through decoupled training and inference, D2 converts the D1 strategy, which depends on online DA3 prior generation, into a student branch that directly predicts task-related structural responses. The purpose is to avoid continuously running DA3 during

final inference while retaining the auxiliary effect of depth-structure priors on detection-feature encoding.

For comparison, we constructed D2-Raw, whose single-channel student head is trained to predict the image-wise normalized DA3 response $\hat{D}$ using a per-element Smooth L1 loss. The prediction, denoted by P_s^{raw} , is transformed through a fixed element-wise mapping into $[P_s^{\text{raw}}, 1 - P_s^{\text{raw}}, (P_s^{\text{raw}})^2]$ to form the three-channel input to PGA. In contrast, D2 directly predicts the three-channel structural prior P_s , with $\{T_d, S_b, R\}$ as the supervision targets. Both variants used the same data split and random seed (2026), with the remaining training settings and validation-based checkpoint-selection rule held constant.

2.4. Non-Rigid Fish Turning–Deformation Coupling Consistency Association

In a tracking-by-detection framework, the detector first provides frame-wise candidate boxes, after which the tracker associates trajectories with detections using cues such as motion prediction, spatial overlap, appearance features, detection confidence, or velocity-direction consistency. For the j -th trajectory and the i -th candidate detection, the baseline association cost is denoted by C_{base} .

Fish exhibit pronounced non-rigid motion. Turning, tail beating, and body bending can contribute to short-term changes in bounding-box area s and aspect ratio r , although detection error, partial occlusion, depth-wise motion, and interactions with neighboring fish can also affect these measurements. If geometric changes are treated independently as abnormalities, normal non-rigid deformation may be misclassified; if geometric constraints are completely relaxed, incorrect identity matching is more likely during fish crossing and reassociation after short-term occlusion.

Therefore, this study introduces Fish Turning–Deformation Coupling Consistency (FTDC), which interprets bounding-box geometric changes conditionally on short-term turning amplitude and incorporates the resulting constraint into the SU-T candidate-matching stage.

2.4.1. Fish Observations and Short-Term Direction Estimation

Let the i -th detection box in frame t be represented by its upper-left and lower-right coordinates (x_1, y_1, x_2, y_2) , with width and height denoted by w and h , respectively. The bounding box is converted into a fish observation vector containing center position, area, and aspect ratio:

$$z_i^t = [c_{x,i}^t, c_{y,i}^t, s_i^t, r_i^t]^\top \quad (35)$$

where the geometric quantities are defined as follows:

$$c_{x,i}^t = \frac{x_{1,i}^t + x_{2,i}^t}{2}, \quad c_{y,i}^t = \frac{y_{1,i}^t + y_{2,i}^t}{2}, \quad s_i^t = w_i^t h_i^t, \quad r_i^t = \frac{w_i^t}{h_i^t} \quad (36)$$

The short-term direction is not used to recover the fish-head orientation or precise posture; instead, it describes the motion trend of a trajectory between adjacent frames. A short-term direction window $k = 1$ is used, and the three consecutive frames, $t - 2$, $t - 1$, and t , form two adjacent single-frame displacement vectors. The historical direction is estimated from the displacement of the trajectory center from $t - 2$ to $t - 1$, while the candidate direction is estimated from the trajectory center in frame $t - 1$ to the center of the current candidate detection:

$$\theta_j^{t-1} = \text{atan2}(c_{y,j}^{t-1} - c_{y,j}^{t-1-k}, c_{x,j}^{t-1} - c_{x,j}^{t-1-k}) \quad (37)$$

$$\theta_{ij}^t = \text{atan2}(c_{y,i}^t - c_{y,j}^{t-1}, c_{x,i}^t - c_{x,j}^{t-1}) \quad (38)$$

To avoid discontinuities when the angle crosses π or $-\pi$, the turning amplitude of a candidate match is defined as follows:

$$\Delta\theta_{ij}^t = \left| \text{wrap}(\theta_{ij}^t - \theta_j^{t-1}) \right| \quad (39)$$

where $\text{wrap}()$ normalizes the angular difference to the range $[-\pi, \pi]$, and the outer absolute-value operation yields a non-negative turning amplitude. When the historical displacement or current candidate displacement is smaller than the motion threshold $\delta_m = 1.0$ pixels, the turning-angle estimate becomes unstable; in this case, the current turning angle is deterministically set to $\Delta\theta_{ij}^t = 0$, and the FTDC penalty is still computed using the envelope intercept under the zero-turning condition. This operation does not unconditionally set the FTDC cost to zero: the additional cost is zero if the area or aspect ratio change does not exceed the corresponding intercept and remains nonzero if it does. The FTDC additional cost is returned directly as zero only when the trajectory history is insufficient, the bounding box is invalid, or FTDC is disabled.

2.4.2. Global Linear Quantile Envelope for Turning–Deformation Coupling

For the j -th trajectory and the i -th candidate detection, logarithmic relative changes are used to describe observed variations in aspect ratio and area:

$$\delta_{ij}^r = \left| \log \frac{r_i^t}{r_j^{t-1}} \right|, \quad \delta_{ij}^s = \left| \log \frac{s_i^t}{s_j^{t-1}} \right| \quad (40)$$

Using logarithmic ratios reduces the influence of target-scale differences and provides more symmetric penalties for increases and decreases in area and aspect ratio.

We hypothesize that larger changes in the trajectory's motion direction may coincide with a wider range of observed bounding-box area and aspect ratio changes. Therefore, a global linear quantile envelope is used to describe the observed range of geometric variation allowed under different turning amplitudes:

$$\delta_{ij}^r \leq a_r \Delta\theta_{ij}^t + b_r, \quad \delta_{ij}^s \leq a_s \Delta\theta_{ij}^t + b_s \quad (41)$$

where a_r and a_s denote the slopes of the aspect ratio and area-change envelopes with increasing turning amplitude, respectively, and b_r and b_s denote the baseline aspect ratio and area fluctuations allowed under the zero-turning condition, respectively. The envelope summarizes variation in projected bounding-box geometry arising from fish motion, imaging, and localization effects.

The envelope parameters are estimated offline using only ground-truth trajectories from the GP+LC training set. Samples are constructed from three strictly consecutive frames, $t - 2$, $t - 1$, and t , of the same trajectory with a fixed interval of one frame. Only annotations with a valid confidence flag ($\text{conf} > 0$) and positive bounding-box width and height are retained. Low-displacement samples are not discarded; instead, their turning angle is set to 0 according to the rule in Section 2.4.1. No additional statistical outlier clipping is performed. Instead, the 95th empirical percentile within each bin is used to estimate the upper boundary of geometric variation, reducing the influence of extreme samples on the fitting results.

The turning-angle range $[0, \pi]$ is divided into 12 equal-width bins, each with a width of $\pi/12$ (15°), and only bins containing at least five samples participate in fitting. The 95th percentiles of the logarithmic area change and logarithmic aspect ratio change are calculated for each valid bin. An unweighted first-order linear fit is then performed using the bin-center angle and the corresponding percentile point to obtain a global linear envelope

rather than a piecewise-linear function. If fewer than two valid bins are available or the fitted slope is negative, the slope is set to zero so that the envelope degenerates to a constant upper bound. The numerical-stability term is set to $\varepsilon = 10^{-6}$.

A total of 19,381 valid three-frame samples were obtained from the GP+LC training set. The fitted envelope coefficients were $a_s = 0.0283$, $b_s = 0.2807$, $a_r = 0.0135$, and $b_r = 0.2621$. These parameters were estimated only once from the training set and were subsequently fixed for validation and testing, without refitting, online updating, or test-data-based parameter adaptation.

The distribution of these training samples, bin-level 95th percentiles, and fitted envelopes are examined in Section 3.7.

2.4.3. FTDC Association Cost and Its Integration into SU-T

For the portion exceeding the reasonable envelope, the FTDC coupling penalty is defined as follows:

$$C_{ij}^{\text{couple}} = \max\left(0, \delta_{ij}^r - a_r \Delta \theta_{ij}^t - b_r\right) + \max\left(0, \delta_{ij}^s - a_s \Delta \theta_{ij}^t - b_s\right) \quad (42)$$

where $\max(0, x)$ denotes the positive-part function. When the turning amplitude is small but the bounding-box area or aspect ratio changes abruptly, the FTDC coupling cost increases, thereby reducing the priority of that candidate match. At larger turning amplitudes, the fitted envelope allows for greater geometric variation before an additional penalty is applied. To prevent the additional cost from excessively perturbing the original association matrix, the FTDC cost is clipped to $[0, 1]$ during inference.

During SU-T trajectory-detection candidate matching, let the original association cost be C_{ij}^{base} . FTDC is then incorporated into the candidate-cost matrix as an additional soft constraint:

$$C_{ij}^{\text{new}} = C_{ij}^{\text{base}} + \lambda_c C_{ij}^{\text{couple}} \quad (43)$$

where the baseline association cost consists of the original SU-T cues, including motion prediction, spatial overlap, and detection confidence. The coupling weight is set by default to $\lambda_c = 0.10$. FTDC is applied only in the first-stage high-confidence candidate association of SU-T and is not enabled in the second-stage low-confidence association. It does not modify the detector, state predictor, trajectory initialization, lost-track retention, termination, or other life-cycle management logic.

2.5. Evaluation Metrics and Experimental Protocol

Detection experiments follow a unified Common Objects in Context (COCO) bounding-box evaluation protocol and report the number of parameters (Params), giga floating-point operations (GFLOPs), average precision over intersection-over-union (IoU) thresholds from 0.50 to 0.95 (AP50:95), average precision at IoU = 0.50 (AP50), average precision at IoU = 0.75 (AP75), and average recall with a maximum of 100 detections per image (AR100). D0, D1, D2, YOLOv8m, YOLO11m, and YOLO26m are evaluated on the same test split. The D2-Raw control uses the same test images and COCO evaluation protocol. Params and GFLOPs are used to compare model size and theoretical computational complexity among the different detection front ends.

The multi-object-tracking experiments use the fixed combined GP+LC MOT test sequences. The main method, overall-method ablation, internal FTDC component ablation, and parameter-sensitivity experiments all use SU-T as the baseline tracker and mainly report multiple-object-tracking accuracy (MOTA), identification F1 score (IDF1), higher-order tracking accuracy (HOTA), identity switches (IDs), and track fragmentations (Frag); false positives (FPs) and false negatives (FNs) are additionally reported in the coupling-

weight sensitivity analysis. The controlled comparison of representative methods uses the same D2 detection results to evaluate SORT, ByteTrack, OC-SORT, SU-T, and SU-T+FTDC under a unified protocol. The FTDC weight-sensitivity experiment uses D2+SU-T and scans five coupling weights: 0.00, 0.05, 0.10, 0.15, and 0.20.

To assess sensitivity to temporal sampling, a paired downsampling experiment was conducted on GP_MOT_004 and LC_MOT_004 at 15 fps and GP_MOT_009 at 10 fps. The sequences at 15 fps were uniformly downsampled by a factor of three, and the sequence at 10 fps by a factor of two, producing a version at 5 fps for each clip. SU-T and SU-T+FTDC were evaluated under both sampling conditions. The overall results were aggregated separately across the same three clips at their native effective frame rates and after downsampling. The comparison reports changes in HOTA, its association accuracy component (AssA), and IDF1 in percentage points, together with the relative change in identity switches.

3. Results

3.1. Implementation Details

The experiments were conducted on a platform equipped with an Intel Core i9-10980XE CPU, two NVIDIA GeForce RTX 4090 GPUs, and 64 GB of memory. The deep learning framework was PyTorch 2.5.1 with CUDA 12.1.

All detectors used the same dataset split, a 120-epoch budget, batch size 32, and evaluation size of 640×640 . RT-DETR used AdamW with learning rates of 4×10^{-4} and 4×10^{-5} for non-backbone and backbone parameters, respectively, weight decay 10^{-4} , and a FlatCosine schedule. Mosaic and MixUp were active during epochs 4–48, and multi-scale training at 480–800 pixels was used before epoch 90. The YOLO models used COCO-pretrained weights and Ultralytics 8.4.75 defaults, with AdamW learning rate 0.002, weight decay 0.0005, linear decay after three warmup epochs, and Mosaic disabled for the final 10 epochs.

All runs completed 120 epochs. RT-DETR evaluated exponential moving average (EMA) weights and reloaded the best first-stage checkpoint at epoch 90. YOLO early stopping with patience 100 was not triggered. Final checkpoints were selected by validation AP50:95 and subsequently evaluated on the test split. D2 was trained with seeds 2026, 42, and 3407 under the same protocol, with AP50:95 reported as mean $\pm$ sample standard deviation.

D0 introduces neither an additional depth-structure prior nor a prior-distillation loss. D1 runs the frozen DA3 frame by frame during both training and inference, generates structural priors from its relative depth responses, and uses them for PGA feature enhancement without an additional depth-prior loss. During training, D2 uses the frozen DA3 to generate teacher structural priors and supervises TinyDepthPriorHead with a Smooth L1 loss; the depth-structure-prior distillation-loss weight is set to $\lambda_{\text{depth}} = 0.01$. None of the three configurations enables the original RT-DETRv4 VFM teacher branch or its distillation loss during downstream training. The DA3 processing resolution is set to 320×320 .

The detection-side comparison includes D0, D1, D2, YOLOv8m, YOLO11m, and YOLO26m, all evaluated on the same test set using the COCO bounding-box protocol. A separate D2-Raw comparison evaluates the simpler depth-prior design under the same protocol. The tracking side uses the same GP+LC MOT test sequences and uniformly reports MOTA, IDF1, HOTA, IDs, and Frag.

All SU-T-based experiments used the same tracker configuration. To accommodate the different effective frame rates, max_time_lost was set to 2 s, corresponding to 10, 20, and 30 frames at 5, 10, and 15 fps, respectively. The FTDC coupling weight was set to $\lambda_c = 0.10$. FTDC was applied only during the first-stage high-confidence association, and

its additional cost was clipped to $[0, 1]$. All other association settings were kept fixed across experiments.

The FTDC envelope-fitting procedure and parameter definitions are described in Section 2.4.2. The envelope is estimated offline only once from the GP+LC training-set ground-truth trajectories using 19,381 valid three-frame samples; the resulting parameters are fixed for the validation and test sets, with no refitting or online updating during inference.

In the tracker comparison, SORT, ByteTrack, OC-SORT, SU-T, and SU-T+FTDC share exactly the same D2 detection results, ground truth, IoU threshold, and evaluation scripts. Parameters use predefined configurations or are determined on the validation set, and the test set is used only for final evaluation.

3.2. Comparison of Detector Complexity and Detection Performance

Table 1 shows that D1 achieves the highest AP50:95 and AR100 (0.857 and 0.903), but the online DA3 increases Params and GFLOPs to 154.55 M and 119.85, respectively. After DA3 is removed during inference, D2 has 19.68 M parameters and 60.90 GFLOPs, only 0.49 M parameters and 4.57 GFLOPs more than D0. Its AP50:95 increases from 0.837 to 0.851, and its AP75 reaches the highest value in the table, 0.949. Compared with YOLOv8m, YOLO11m, and YOLO26m, the AP50:95 of D2 is higher by 0.029, 0.022, and 0.035, respectively, indicating that D2 retains the detection gains provided by the depth-structure prior with only a small additional computational cost.

Table 1. Comparison of model complexity and detection performance among different detectors (test split).

Detector	Params (M)	GFLOPs	AP50:95	AP50	AP75	AR100
D0	19.19	56.33	0.837	0.971	0.937	0.888
D1	154.55	119.85	0.857	0.980	0.943	0.903
D2	19.68	60.90	0.851	0.979	0.949	0.893
YOLOv8m	25.84	78.70	0.822	0.969	0.906	0.854
YOLO11m	20.03	67.60	0.829	0.963	0.898	0.859
YOLO26m	20.35	67.80	0.816	0.949	0.892	0.851

Across the three training runs, D2 achieved a test AP50:95 of 0.8507 ± 0.0015 (mean $\pm$ sample standard deviation; $n = 3$), corresponding to a standard deviation of 0.15 percentage points. The D2 values in Tables 1 and 2 correspond to seed 2026.

Table 2 compares D2-Raw with D2 under the same seed and training protocol. D2-Raw achieved AP50:95 of 0.842, whereas D2 achieved 0.851; the corresponding AP75 values were 0.914 and 0.949. AP50 increased by 0.006 with D2, while AR100 changed by 0.001.

Table 2. Comparison of the simple depth-prior control and mapped-prior D2 on the GP+LC test split (1350 images; seed 2026).

Variant	AP50:95	AP50	AP75	AR100
D2-Raw	0.842	0.973	0.914	0.892
D2	0.851	0.979	0.949	0.893

3.3. Ablation Study of PGA Components

To analyze the independent contributions of multi-scale feature gating and attention bias within PGA, component ablation is performed on the online frozen-DA3 prior pathway. Except for D0, the other three configurations use exactly the same online DA3 structural prior and differ only in whether feature gating and attention bias are enabled. When

both prior-injection mechanisms are disabled, the DA3 prior does not enter the detection features; therefore, D0 is used as the baseline. The results are shown in Table 3.

Table 3. Ablation results of different PGA components.

Variant	Online DA3 Prior	Feature Gating	Attention Bias	AP50:95	AP50	AP75	AR100
Baseline (D0)	×	×	×	0.837	0.971	0.937	0.888
Gate only	✓	✓	×	0.848	0.972	0.945	0.895
Attention only	✓	×	✓	0.852	0.976	0.935	0.898
Full PGA (D1)	✓	✓	✓	0.857	0.980	0.943	0.903

A checkmark (✓) indicates enabled; a cross (×) indicates disabled. Bold values indicate the best result in each metric.

Table 3 shows that Gate only increases AP50:95 and AP75 to 0.848 and 0.945, respectively, while Attention only achieves AP50:95 and AR100 values of 0.852 and 0.898. Full PGA achieves the highest AP50:95, AP50, and AR100 values (0.857, 0.980, and 0.903), with an AP75 of 0.943. The two mechanisms emphasize gains in different metrics, and their joint use provides the best overall performance, indicating that feature gating and attention bias are complementary.

3.4. Detection Feature Visualization Analysis

To inspect spatial network responses [40], class-agnostic activation maps were generated for D0 and D2 on five representative images (Figure 4). A forward hook captured the HybridEncoder (encoder) output, from which the final P5 tensor (stride 32,256 channels) was selected. At each spatial location, the mean absolute activation was computed across channels. The resulting map was bilinearly resized to the input image size, min–max-normalized within each image, displayed with the JET color map, and overlaid on the input with an opacity of 0.45.

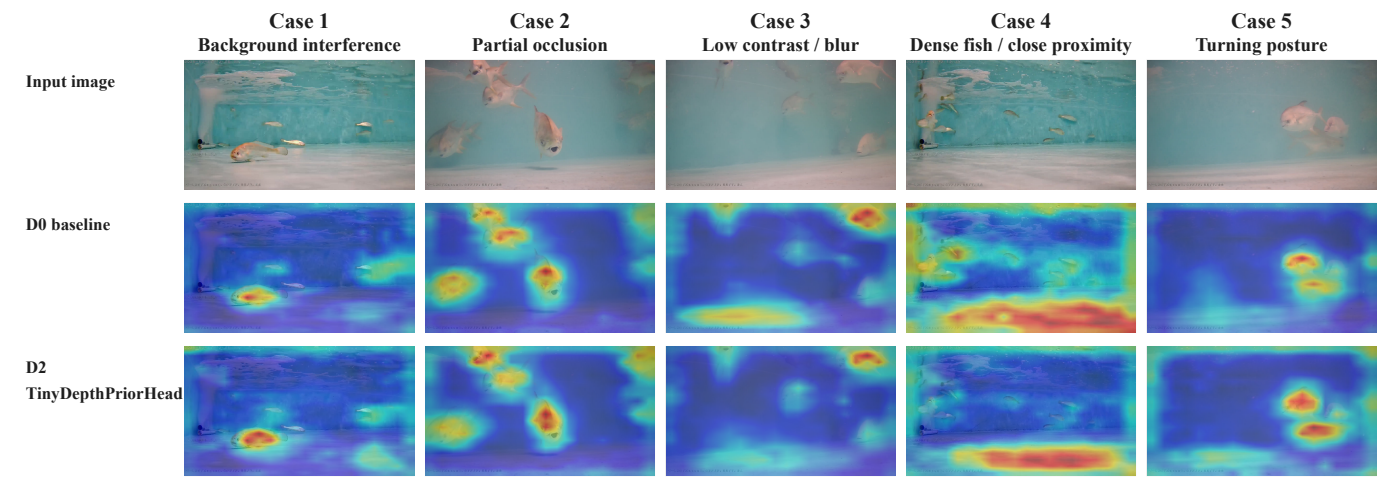


Figure 4. Class-agnostic P5 feature-activation maps for D0 and D2, computed from the stride-32 HybridEncoder output using the channel-wise mean of absolute activations. Colors show response magnitudes normalized within each image.

In the displayed examples, D2’s higher-response regions appear more localized around fish bodies, whereas D0 exhibits more diffuse spatial responses.

3.5. Comparison with Representative Multi-Object-Tracking Methods

To evaluate the relative performance of the final tracking configuration, SORT, ByteTrack, OC-SORT, SU-T, and SU-T+FTDC are compared using the same D2 detection input, ground truth, IoU threshold, and evaluation scripts. The results are shown in Table 4.

Table 4 shows that the MOTA values of SORT, ByteTrack, OC-SORT, and SU-T are 84.2%, 86.5%, 87.1%, and 87.6%, respectively. After FTDC is introduced, SU-T achieves 88.3% MOTA, 77.4% IDF1, and 62.8% HOTA, while IDs and Frag decrease to 60 and 78. Compared with OC-SORT, MOTA, IDF1, and HOTA improve by 1.2, 1.4, and 1.4 percentage points, respectively. Compared with the original SU-T, IDs decrease from 66 to 60 and Frag from 86 to 78, indicating that FTDC mainly improves identity preservation and trajectory continuity.

Table 4. Comparison of representative trackers under the same D2 detection input.

Tracker	FTDC	MOTA	IDF1	HOTA	IDs	Frag
SORT	×	84.2%	70.5%	57.5%	158	185
ByteTrack	×	86.5%	75.0%	60.8%	88	112
OC-SORT	×	87.1%	76.0%	61.4%	72	92
SU-T	×	87.6%	76.8%	62.1%	66	86
SU-T	✓	88.3%	77.4%	62.8%	60	78

A checkmark (✓) indicates enabled; a cross (×) indicates disabled. Bold values indicate the best result in each metric.

3.6. Ablation Study of the Overall Method

To verify the independent contributions and combined effect of the two core designs, D0+SU-T is used as the baseline, and D2 and FTDC are added separately. The results are shown in Table 5.

Table 5. Ablation results of the overall method.

Depth-Structure Prior Enhancement	FTDC	MOTA	IDF1	HOTA	IDs	Frag
×	×	86.7%	75.6%	60.9%	90	117
✓	×	87.6%	76.8%	62.1%	66	86
×	✓	87.2%	76.0%	61.4%	85	111
✓	✓	88.3%	77.4%	62.8%	60	78

A checkmark (✓) indicates enabled; a cross (×) indicates disabled. Bold values indicate the best result in each metric.

With D2 alone, MOTA, IDF1, and HOTA increase from 86.7%, 75.6%, and 60.9% to 87.6%, 76.8%, and 62.1%, respectively, while IDs/Frag decrease from 90/117 to 66/86. With FTDC alone, the corresponding metrics reach 87.2%, 76.0%, and 61.4%, and IDs/Frag decrease to 85/111. When both improvements are enabled, MOTA, IDF1, and HOTA reach 88.3%, 77.4%, and 62.8%, respectively, while IDs/Frag further decrease to 60/78. These results indicate that detection enhancement and association constraints can contribute independently and provide complementary benefits when used together.

3.7. Training-Set Turning–Geometry Relationship

Figure 5 plots short-term trajectory-direction change against the absolute logarithmic changes in bounding-box area and aspect ratio for the 19,381 valid training-set samples described in Section 2.4.2. Each of the 12 angle bins contained at least 770 samples. The 1418 low-displacement samples were assigned a turning angle of zero under the FTDC fitting rule.

The fitted upper boundary rises with turning magnitude for both area and aspect ratio. The bin-level fit is stronger for area change ($R^2_{\text{bin}} = 0.833$) than for aspect ratio change ($R^2_{\text{bin}} = 0.536$). Across individual samples, the rank correlations are positive but modest

($\rho_{\text{raw}} = 0.259$ and 0.296 , respectively). The envelope captures an increase in the upper range of geometric variation, with substantial scatter among individual observations.

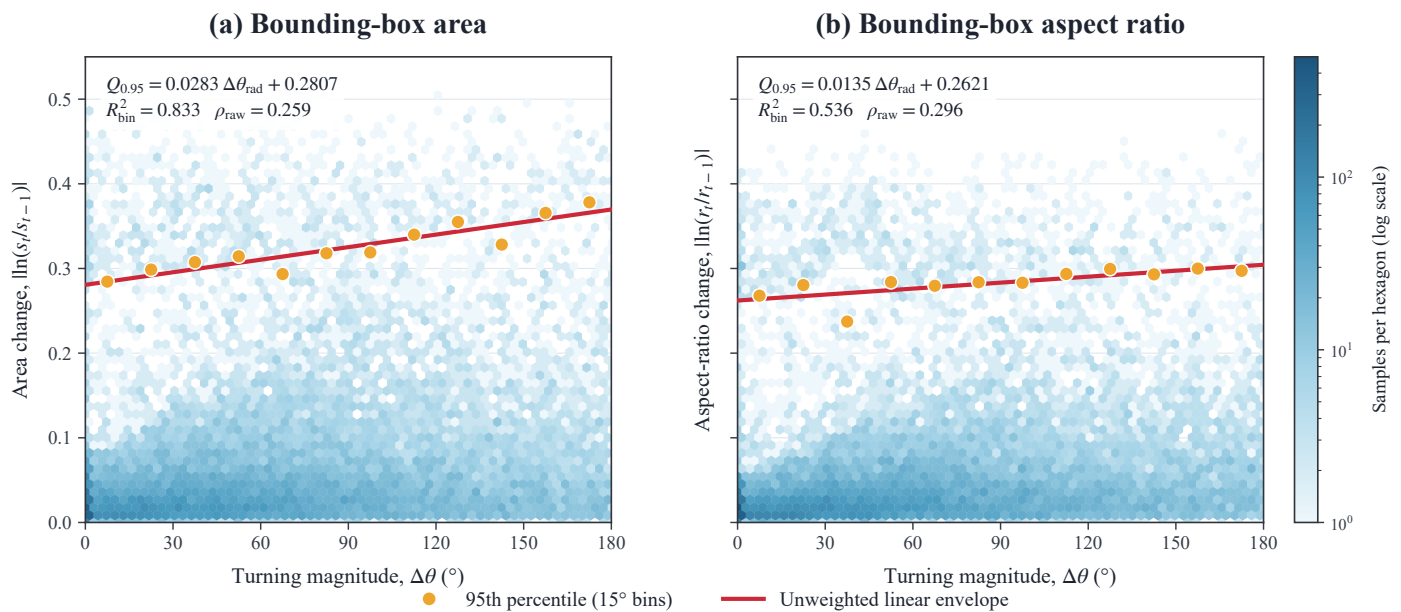


Figure 5. Training-set relationship between short-term trajectory-direction change and bounding-box variation. Blue hexagons show sample density; orange circles show the 95th percentile in each 15° bin; red lines are unweighted linear fits to the 12 percentile points. The horizontal axis is displayed in degrees, while the fitted slopes use radians. R^2_{bin} evaluates the fit to the bin-level percentiles, and ρ_{raw} is Spearman’s rank correlation across individual samples.

3.8. FTDC Component Ablation and Parameter Sensitivity Analysis

With D2+SU-T and the evaluation protocol fixed, only the internal components of FTDC are varied. Area only uses area variation, Ratio only uses aspect ratio variation, Geometry jointly uses both geometric cues without introducing the turning condition, and the complete FTDC uses the turning–deformation coupling envelope. The results are shown in Table 6.

The MOTA values of Area only and Ratio only are 87.9% and 87.8%, respectively. Geometry further reaches 88.1% MOTA, 77.3% IDF1, and 62.6% HOTA, with IDs/Frag of 62/80. The complete FTDC reaches 88.3%, 77.4%, and 62.8%, while IDs/Frag decrease to 60/78. Therefore, area and aspect ratio provide complementary information, and turning-conditioned coupling yields additional association gains over unconditional geometric constraints.

The effect of the coupling weight on association performance is further examined, and the sensitivity analysis is shown in Table 7.

Table 6. Ablation results of FTDC components (D2+SU-T).

Variant	Turning	Area	Aspect Ratio	Coupling Envelope	MOTA	IDF1	HOTA	IDs	Frag
SU-T	×	×	×	×	87.6%	76.8%	62.1%	66	86
Area only	×	✓	×	×	87.9%	77.1%	62.4%	63	82
Ratio only	×	×	✓	×	87.8%	77.0%	62.3%	64	83
Geometry	×	✓	✓	×	88.1%	77.3%	62.6%	62	80
FTDC	✓	✓	✓	✓	88.3%	77.4%	62.8%	60	78

A checkmark (✓) indicates enabled; a cross (×) indicates disabled. Boldface highlights the complete FTDC configuration.

Table 7. Sensitivity analysis of the FTDC coupling weight (D2+SU-T).

λ_c	MOTA	IDF1	HOTA	IDs	Frag	FP	FN
0.00	87.6%	76.8%	62.1%	66	86	657	395
0.05	87.8%	77.1%	62.3%	63	82	647	387
0.10	88.3%	77.4%	62.8%	60	78	608	388
0.15	88.1%	77.2%	62.6%	62	80	618	390
0.20	87.5%	76.5%	61.9%	73	95	656	395

Table 7 shows that when λ_c increases from 0.00 to 0.10, MOTA, IDF1, and HOTA increase from 87.6%, 76.8%, and 62.1% to 88.3%, 77.4%, and 62.8%, respectively, while IDs and Frag decrease from 66 and 86 to 60 and 78. Performance declines when the weight is increased further; at $\lambda_c = 0.20$, IDs and Frag rise to 73 and 95, indicating that an excessively strong penalty suppresses reasonable non-rigid variation and instead increases association fragmentation.

Therefore, $\lambda_c = 0.10$ is used as the default FTDC weight.

3.9. Frame-Rate Sensitivity Analysis

At the native effective frame rates, adding FTDC to SU-T increased aggregate HOTA, AssA, and IDF1 by 0.7, 1.1, and 0.6 percentage points, respectively (Table 8). Identity switches decreased from 59 to 53, a relative reduction of 10.2%. After downsampling the same clips to 5 fps, the aggregate gains decreased to 0.2 percentage points in HOTA, 0.4 in AssA, and 0.1 in IDF1. Identity switches decreased from 120 to 115, a relative reduction of 4.2%.

Table 8. Changes in tracking performance after adding FTDC to SU-T at native effective frame rates and after paired downsampling to 5 fps.

Sequence	Sampling	Δ HOTA (pp)	Δ AssA (pp)	Δ IDF1 (pp)	IDSW Change
GP_MOT_004	15 fps	+0.7	+1.1	+0.6	−10.0%
GP_MOT_004	15 → 5 fps	+0.2	+0.5	+0.1	−5.8%
GP_MOT_009	10 fps	+0.6	+0.9	+0.5	−8.7%
GP_MOT_009	10 → 5 fps	−0.1	+0.2	−0.1	−1.9%
LC_MOT_004	15 fps	+0.8	+1.2	+0.7	−11.5%
LC_MOT_004	15 → 5 fps	+0.3	+0.6	+0.2	−6.3%
Overall	Native	+0.7	+1.1	+0.6	−10.2%
Overall	Downsampled	+0.2	+0.4	+0.1	−4.2%

Δ HOTA, Δ AssA, and Δ IDF1 are percentage-point differences between SU-T+FTDC and SU-T. IDSW change is relative to SU-T. The overall results combine the three paired clips separately under each sampling condition. IDSW decreased from 59 to 53 at the native frame rates and from 120 to 115 after downsampling. Boldface marks the pooled results for the three paired clips.

The sequence-level results showed smaller gains after downsampling. GP_MOT_004 and LC_MOT_004 retained positive changes in HOTA and IDF1 at 5 fps, whereas GP_MOT_009 showed decreases of 0.1 percentage points in both metrics despite a small increase in AssA. FTDC therefore retained a positive aggregate benefit at the lower sampling rate, but the improvement was smaller and varied across sequences.

3.10. Visualization Analysis of Multi-Object-Tracking Results

To visually demonstrate the effect of FTDC on identity preservation, three representative scenarios involving fish crossing, close-range overlap, or reassociation after short-term occlusion are selected for comparison, as shown in Figure 6. Figure 6 compares D2+SU-T with D2+SU-T+FTDC ($\lambda_c = 0.10$).



Figure 6. Identity preservation comparison before and after introducing FTDC in fish crossing and reassociation after short-term occlusion. The red label on the left indicates an ID switch with Original SU-T, whereas the green label indicates identity preservation with SU-T+FTDC. Labels beneath each panel repeat the displayed identities for fish A and B.

The examples in Figure 6 show that Original SU-T exhibits ID swaps during reassociation after fish crossing, close-range overlap, or short-term occlusion, whereas SU-T+FTDC preserves the original identities in the same clips. Together with the reductions in IDs and Frag shown in Table 4, these examples indicate that the turning–deformation coupling soft constraint of FTDC can suppress candidate matches that are clearly inconsistent with recent motion–deformation relationships in these representative difficult segments.

4. Discussion

4.1. Effects of Depth-Structure Priors on Underwater Fish Detection

The detector comparison in Table 1 indicates that incorporating relative depth structure improves fish detection relative to the RGB baseline. Underwater scattering and attenuation weaken appearance cues [3,38,39]. Relative depth can provide complementary information about target–background structure under these conditions.

D2 uses three structural cues derived from the same normalized relative-depth response. It achieved higher AP50:95 and AP75 than D2-Raw in the seed-2026 run (Table 2). The two variants differ in both the student-head output dimension and the fixed prior transform; so, the comparison concerns the overall prior design.

The PGA ablation in Table 3 further indicates that feature gating and attention bias play complementary roles. Feature gating adjusts the local response magnitude of multi-scale visual features, whereas attention bias modifies the spatial relationships modeled within the Transformer encoder. The benefit of using both mechanisms suggests that the prior contributes at both the regional-feature and spatial-interaction levels.

Table 1 also reveals a practical trade-off between D1 and D2. Online DA3 inference provides the strongest access to relative-depth information but substantially increases model size and computation. D2 instead transfers task-related structural responses from the frozen DA3 teacher to TinyDepthPriorHead and retains most of the detection benefit

without continuously running the depth foundation model. D1 therefore serves mainly to verify the value of online structural information, whereas D2 is the practical inference configuration used in the final tracking pipeline.

4.2. Effects of Turning–Deformation Coupling on Fish Identity Association

The four-way ablation in Table 5 separates the detection and association contributions. With SU-T fixed, replacing D0 with D2 increases HOTA from 60.9% to 62.1%. Adding FTDC raises HOTA from 60.9% to 61.4% with D0 and from 62.1% to 62.8% with D2. The depth-prior branch affects the candidate observations available for association, while FTDC adjusts their matching cost. Under partial occlusion, improved localization may help preserve usable observations; during turning, FTDC accommodates larger geometric changes within the fitted envelope. Its effect under occlusion therefore depends on the availability of suitable detections.

Recent fish trackers address underwater association through detection correction and Re-ID in FSTA [20], memory aggregation in UMOTMA [22], and stronger appearance embeddings in improved DeepSORT [21]. Other methods explore nonlinear motion, spatiotemporal fusion, scale variation, and sequence modeling [5,13,23–28]. SU-T provides unscented motion prediction and FishIoU geometry matching [5]. FTDC supplements these mechanisms with an empirical link between trajectory-direction change and allowable area and aspect ratio variation, adding a soft cost in the first association stage without an additional Re-ID branch.

Figure 5 shows a rising upper boundary of bounding-box variation with trajectory-direction change in the training data. The area trend is clearer than the aspect ratio trend, while the modest sample-level correlations indicate substantial variation around both trends. These patterns support an empirical tolerance envelope for association. The figure summarizes the training samples used to fit the envelope; its stability across independent sequences requires further evaluation.

The component ablation in Table 6 shows that area and aspect ratio changes each provide useful association information, and their combination offers a stronger geometric description. The complete FTDC performs better than the unconditional geometry variant, although the additional tracking gain is modest.

The coupling-weight analysis in Table 7 further characterizes FTDC as an auxiliary soft constraint. Moderate weighting improves association, whereas excessive weighting can reject valid candidates during large but valid bounding-box changes and thereby reduce trajectory continuity. FTDC is therefore most effective when it supplements, rather than replaces, the original motion and spatial association mechanisms of SU-T.

The paired downsampling results in Table 8 indicate sensitivity to the temporal sampling interval. FTDC estimates turning from adjacent displacement vectors, whose physical duration increases as the effective frame rate decreases. The fitted envelope is therefore applied to observations spanning different time intervals. In the evaluated clips, the association benefit became smaller at 5 fps, and one sequence showed a slight decrease in HOTA and IDF1. Direction estimates over a fixed physical time window warrant evaluation to determine whether this sensitivity can be reduced.

The qualitative cases in Figure 6 provide complementary evidence. During fish crossing, close-range overlap, and reassociation after short-term occlusion, the original SU-T produces identity exchanges in the selected examples, whereas SU-T+FTDC preserves the original identities. These cases are consistent with the controlled association results in Tables 4–6. Because all trackers use the same D2 detections in Table 4 and the proposed pipeline introduces no additional Re-ID branch, the observed differences mainly reflect the association strategy based on motion and bounding-box geometry rather than changes in

detection or additional appearance embeddings. FTDC reduces, but does not eliminate, identity switches and trajectory fragmentation (Table 4).

4.3. Implications for Continuous Fish Monitoring in Marine Aquaculture

The experiments address continuous visual observation of golden pompano and large yellow croaker in a land-based aquaculture tank. In this setting, underwater image degradation affects whether fish are detected and accurately localized, while crossing and non-rigid motion complicate the linking of successive observations to the same individual. The depth-structure-prior-enhanced detector and FTDC address these complementary sources of uncertainty. Their combined contribution is therefore a more consistent record of fish positions and identities within the observed video sequences.

Identity continuity is relevant to aquaculture applications that use trajectories to characterize fish movement or behavior [1,2,13]. An identity switch can join observations from different fish, whereas fragmentation interrupts the temporal record of one fish; either can complicate subsequent trajectory analysis. Under identical D2 detections, adding FTDC to SU-T reduced identity switches from 66 to 60 and fragmentations from 86 to 78, with corresponding improvements in IDF1 and HOTA (Table 4). These results support improved association in the evaluated sequences and complement the frame-level detection gains.

The resulting trajectories provide a basis for movement and activity analysis, although their value for specific behavioral or management tasks requires separate validation.

For practical video monitoring, D2 uses TinyDepthPriorHead at inference, avoiding frame-wise DA3 processing. On an RTX 4090, the complete D2+SU-T+FTDC pipeline achieved an average throughput of 16.4 FPS (compared with 17.0 FPS for D2+SU-T) over three runs on the 1350 test frames, exceeding the highest evaluation sampling rate of 15 fps. Adding FTDC increased processing time by approximately 4% relative to D2+SU-T. These measurements characterize average throughput on the evaluated test clips, rather than sustained processing of the original 30 fps input.

4.4. Limitations and Future Work

Several limitations should be considered when interpreting the present results. First, FTDC is currently integrated only into SU-T. Although its formulation is based on candidate-level motion and bounding-box geometry and is therefore conceptually compatible with other tracking-by-detection frameworks, its transferability to different baseline trackers has not yet been experimentally demonstrated. Future work will evaluate FTDC in additional association frameworks to determine whether the observed improvements are specific to SU-T or can be generalized to other online trackers.

Second, the dataset combines effective frame rates of 5, 10, and 15 fps, selected according to sequence-level motion and redundancy. Frame rate and scene dynamics may therefore be confounded. The three-clip paired experiment showed a smaller aggregate FTDC benefit at 5 fps, with variation among sequences, consistent with earlier fish-tracking studies reporting performance differences across frame rates [34]. Additional clips spanning different motion conditions are needed to characterize this dependence.

Third, the experimental dataset contains two marine aquaculture species, golden pompano and large yellow croaker, collected in a land-based aquaculture environment. The test clips came from an original video separate from the training and validation source, but both videos were recorded at the same site on the same day. Training and validation also shared the morning source video; so, validation performance may reflect source-specific recording conditions. The results therefore do not establish generalization to other recording days, other species, substantially different stocking densities, offshore cages, or water bodies with different optical characteristics. In addition, the structural

priors derived from DA3 are physics-inspired task priors rather than calibrated physical measurements. Water clarity, illumination, and imaging distance may affect D2's learned structural responses; changes in species or swimming dynamics may also affect the fitted FTDC envelope. Both components require validation in the target environment.

Fourth, the repeated-seed analysis characterized variation in D2 AP50:95. Variability in the comparative detection gains and downstream tracking performance requires additional repeated experiments.

Finally, the current study focuses on detection and multi-object-tracking performance and does not directly evaluate downstream applications based on the resulting trajectories. Future research may integrate the proposed tracking framework with fish counting, behavior recognition, movement analysis, or other longitudinal monitoring tasks to determine whether improvements in trajectory continuity translate into measurable benefits for practical aquaculture observation.

5. Conclusions

To address underwater imaging degradation and non-rigid fish motion in marine aquaculture fish multi-object tracking, this study proposes depth-structure-prior-enhanced detection and FTDC turning–deformation coupling association. D2 uses DA3 to supervise TinyDepthPriorHead in learning task-related structural priors and does not require online DA3 inference, while FTDC introduces a coupling constraint between short-term turning and changes in area and aspect ratio during the first-stage high-confidence matching of SU-T.

D2 increases AP50:95 from 0.837 to 0.851 and achieves an AP75 of 0.949, with 19.68 M parameters and 60.90 GFLOPs. Under the same D2 detection input, D2+SU-T+FTDC achieves 88.3% MOTA, 77.4% IDF1, and 62.8% HOTA, while IDs and Frag decrease to 60 and 78, outperforming SORT, ByteTrack, OC-SORT, and the original SU-T. The overall ablation study demonstrates that the two improvements contribute independently and achieve the best overall performance when combined.

Future work will further evaluate the transferability of FTDC to other baseline trackers and test short-term direction estimates over a fixed physical time window across frame rates and motion speeds. The generalization of the depth-structure prior and turning–deformation coupling association will also be evaluated across additional marine aquaculture fish species, stocking densities, and more complex water conditions.

Author Contributions: Conceptualization, Z.L. and S.Z.; methodology, Z.L. and F.Y.; software, F.Y.; validation, F.Y., G.H., L.Q. and Y.Y.; formal analysis, Z.L. and F.Y.; investigation, F.Y., G.H., L.Q. and Y.Y.; resources, Z.L., Y.Y. and S.Z.; data curation, F.Y., G.H. and L.Q.; writing—original draft preparation, Z.L. and F.Y.; writing—review and editing, Z.L., F.Y. and S.Z.; visualization, F.Y.; supervision, Z.L. and S.Z.; project administration, Z.L. and S.Z.; funding acquisition, Z.L. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2022YFD2401304; the Scientific Research Program of the Department of Education of Guangdong Province, grant numbers 2024ZDZX4060 and 2025KTSCX047; and the Zhanjiang Science and Technology Program, grant number 2024B01067.

Institutional Review Board Statement: Ethical approval is waived due to this study not involving animal experiments.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request. The data are not publicly available because they are being used in other ongoing studies.

Acknowledgments: The authors thank the staff of the Waipo Base of Guangdong Hetai Fishery Technology Co., Ltd., for providing the experimental site and assistance with underwater video collection. During the preparation of this manuscript, the authors used ChatGPT (GPT-5.6 sol) for English translation and language editing. The authors reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Cui, M.; Liu, X.; Liu, H.; Zhao, J.; Li, D.; Wang, W. Fish tracking, counting, and behaviour analysis in digital aquaculture: A comprehensive survey. *Rev. Aquac.* **2025**, *17*, e13001. <https://doi.org/10.1111/raq.13001>.
2. Liu, C.; Wang, Z.; Li, Y.; Zhang, Z.; Li, J.; Xu, C.; Du, R.; Li, D.; Duan, Q. Research progress of computer vision technology in abnormal fish detection. *Aquac. Eng.* **2023**, *103*, 102350. <https://doi.org/10.1016/j.aquaeng.2023.102350>.
3. Hao, Y.; Yuan, Y.; Zhang, H.; Zhang, Z. Underwater optical imaging: Methods, applications and perspectives. *Remote Sens.* **2024**, *16*, 3773. <https://doi.org/10.3390/rs16203773>.
4. Dawkins, M.; Prior, J.; Lewis, B.; Faillettaz, R.; Banez, T.; Salvi, M.; Rollo, A.; Simon, J.; Campbell, M.; Lucero, M.; et al. FishTrack23: An ensemble underwater dataset for multi-object tracking. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 7152–7161. <https://doi.org/10.1109/WACV57701.2024.00701>.
5. Li, W.; Liu, Y.; Guo, Q.; Wei, Y.; Leo, H.L.; Li, Z. When trackers date fish: A benchmark and framework for underwater multiple fish tracking. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 6459–6467. <https://doi.org/10.1609/aaai.v40i8.37574>.
6. Jian, M.; Yang, N.; Tao, C.; Zhi, H.; Luo, H. Underwater object detection and datasets: A survey. *Intell. Mar. Technol. Syst.* **2024**, *2*, 9. <https://doi.org/10.1007/s44295-024-00023-6>.
7. Yu, H.; Yao, X.; Zhang, R.; Liang, X.; Li, R.; Li, X. Marine aquaculture biological object detection based on color-line image enhancement and improved YOLOv7 model. *Comput. Electron. Agric.* **2025**, *239*, 111119. <https://doi.org/10.1016/j.compag.2025.111119>.
8. Liu, L.; Chu, C.; Chen, C.; Huang, S. MarineYOLO: Innovative deep learning method for small target detection in underwater environments. *Alex. Eng. J.* **2024**, *104*, 423–433. <https://doi.org/10.1016/j.aej.2024.07.126>.
9. Vijayalakshmi, M.; Sasithradevi, A. AquaYOLO: Advanced YOLO-based fish detection for optimized aquaculture pond monitoring. *Sci. Rep.* **2025**, *15*, 6151. <https://doi.org/10.1038/s41598-025-89611-y>.
10. Silva-Alvarado, V.L.; Ahmad, A.; Sendra, S.; Lloret, J. Development of a novel convolution to interactive capture and recalibration enhancement module for underwater fish detection in sensor networks. *Sensors* **2026**, *26*, 4290. <https://doi.org/10.3390/s26134290>.
11. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. <https://doi.org/10.1109/CVPR52733.2024.01605>.
12. Lin, H.; Chen, S.; Liew, J.H.; Chen, D.Y.; Li, Z.; Zhao, Y.; Peng, S.; Guo, H.; Zhou, X.; Shi, G.; et al. Depth Anything 3: Recovering the visual space from any views. In Proceedings of the International Conference on Learning Representations (ICLR), Rio de Janeiro, Brazil, 23–27 April 2026. Available online: <https://openreview.net/forum?id=yirunib818> (accessed on 3 September 2026).
13. Mei, Y.; Sun, B.; Li, D.; Yu, H.; Qin, H.; Liu, H.; Yan, N.; Chen, Y. Recent advances of target tracking applications in aquaculture with emphasis on fish. *Comput. Electron. Agric.* **2022**, *201*, 107335. <https://doi.org/10.1016/j.compag.2022.107335>.
14. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Kim, T.K. Multiple object tracking: A literature review. *Artif. Intell.* **2021**, *293*, 103448. <https://doi.org/10.1016/j.artint.2020.103448>.
15. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468. <https://doi.org/10.1109/ICIP.2016.7533003>.
16. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649. <https://doi.org/10.1109/ICIP.2017.8296962>.
17. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-object tracking by associating every detection box. In *Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 1–21. https://doi.org/10.1007/978-3-031-20047-2_1.
18. Cao, J.; Pang, J.; Weng, X.; Khirrodar, R.; Kitani, K. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 9686–9696. <https://doi.org/10.1109/CVPR52729.2023.00934>.

19. Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv* **2022**, arXiv:2206.14651. <https://doi.org/10.48550/arXiv.2206.14651>.
20. Liu, T.; He, S.; Liu, H.; Gu, Y.; Li, P. A robust underwater multiclass fish-school tracking algorithm. *Remote Sens.* **2022**, *14*, 4106. <https://doi.org/10.3390/rs14164106>.
21. Liu, S.; Zhang, J.; Zheng, H.; Qian, C.; Liu, S. An improved DeepSORT-based model for multi-target tracking of underwater fish. *J. Mar. Sci. Eng.* **2025**, *13*, 1256. <https://doi.org/10.3390/jmse13071256>.
22. Hao, Z.; Qiu, J.; Zhang, H.; Ren, G.; Liu, C. UMOTMA: Underwater multiple object tracking with memory aggregation. *Front. Mar. Sci.* **2022**, *9*, 1071618. <https://doi.org/10.3389/fmars.2022.1071618>.
23. Yu, J.; Liu, L.; Xu, L.; Yu, H.; Chen, Y. Individual motion feature extraction method for sea bass based on improved YOLOv8 and ByteTrack. *Trans. Chin. Soc. Agric. Eng.* **2025**, *41*, 182–190. <https://doi.org/10.11975/j.issn.1002-6819.202409089>.
24. Li, W.; Liu, Y.; Wang, W.; Li, Z.; Yue, J. TFMFT: Transformer-based multiple fish tracking. *Comput. Electron. Agric.* **2024**, *217*, 108600. <https://doi.org/10.1016/j.compag.2023.108600>.
25. Liu, Y.; Li, B.; Zhou, X.; Li, D.; Duan, Q. FishTrack: Multi-object tracking method for fish using spatiotemporal information fusion. *Expert Syst. Appl.* **2024**, *238*, 122194. <https://doi.org/10.1016/j.eswa.2023.122194>.
26. Li, Y.; Han, L. A spatiotemporal transformer with cross-frame encoding and trajectory-aware decoding for multi-target fish tracking. *Sci. Rep.* **2026**, *16*, 1997. <https://doi.org/10.1038/s41598-025-31686-8>.
27. Yao, M.; Huo, Y.; Tian, Q.; Zhao, J.; Liu, X.; Wang, R.; Xue, L.; Wang, H. FMRFT: Fusion Mamba and DETR for Query Time Sequence Intersection Fish Tracking. *Comput. Electron. Agric.* **2025**, *237*, 110742. <https://doi.org/10.1016/j.compag.2025.110742>.
28. Høgstædt, E.U.; Schellewald, C.; Stahl, A.; Mester, R. A multi-purpose tracking framework for salmon welfare monitoring in challenging environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Honolulu, HI, USA, 19–20 October 2025; pp. 2153–2162. <https://doi.org/10.1109/ICCVW69036.2025.00225>.
29. Hou, R.; Xu, B.; Ren, T.; Wu, G. MTNet: Learning modality-aware representation with Transformer for RGBT tracking. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 10–14 July 2023; pp. 1163–1168. <https://doi.org/10.1109/ICME55011.2023.00203>.
30. Xu, B.; Hou, R.; Ren, T.; Zhou, D.; Wu, G.; Cao, J. Learning frequency and memory-aware prompts for multi-modal object tracking. *Pattern Recognit.* **2026**, *179*, 113532. <https://doi.org/10.1016/j.patcog.2026.113532>.
31. Xu, B.; Fang, Y.; Hou, R.; Ren, T. STFormer: RGB-T tracking via spatial-temporal interaction Transformer. *Image Vis. Comput.* **2026**, *168*, 105929. <https://doi.org/10.1016/j.imavis.2026.105929>.
32. Li, Y.; Tian, Y.; Quan, J.; Sui, J. Angle-aware multi-object fish tracking with unique forward orientation estimation. *Aquac. Eng.* **2026**, *112*, 102658. <https://doi.org/10.1016/j.aquaeng.2025.102658>.
33. Abangan, A.S.; Bürgi, K.; Méhault, S.; Deroine, M.; Kopp, D.; Faillettaz, R. Assessment of sustainable baits for passive fishing gears through automatic fish behavior recognition. *Sci. Rep.* **2024**, *14*, 13110. <https://doi.org/10.1038/s41598-024-63929-5>.
34. Kandimalla, V.; Richard, M.; Smith, F.; Quirion, J.; Torgo, L.; Whidden, C. Automated detection, classification and counting of fish in fish passages with deep learning. *Front. Mar. Sci.* **2022**, *8*, 823173. <https://doi.org/10.3389/fmars.2021.823173>.
35. Liao, Z.; Zhao, Y.; Shan, X.; Yan, Y.; Liu, C.; Lu, L.; Ji, X.; Chen, J. RT-DETRv4: Painlessly furthering real-time object detection with vision foundation models. *arXiv* **2025**, arXiv:2510.25257. <https://doi.org/10.48550/arXiv.2510.25257>.
36. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531. <https://doi.org/10.48550/arXiv.1503.02531>.
37. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. <https://doi.org/10.1007/s11263-021-01453-z>.
38. Akkaynak, D.; Treibitz, T. A revised underwater image formation model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6723–6732. <https://doi.org/10.1109/CVPR.2018.00703>.
39. Akkaynak, D.; Treibitz, T. Sea-Thru: A method for removing water from underwater images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1682–1691. <https://doi.org/10.1109/CVPR.2019.00178>.
40. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014*; Springer: Cham, Switzerland, 2014; pp. 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.