香 港 大 學
THE UNIVERSITY OF HONG KONG

NEURAL INFORMATION PROCESSING SYSTEMS
Sydney 2026

TrackFish3D

Self-Supervised 3D Tracking of Schooling Fish from Multi-view Videos

Patt Phurtivilai[1] · **Zhiyang (Frank) Dou**[1] · **Yifan Wu**[1] · **Kinfung Chu**[4]
Yuan Liu[2] · **Lei Yang**[1] · **Wenping Wang**[3] · **Taku Komura**[1]

[1] The University of Hong Kong, Hong Kong, China
[2] Hong Kong University of Science and Technology, Hong Kong, China
[3] Texas A&M University, College Station, TX, US
[4] Centre for Transformative Garment Production

Self-supervised multi-view association and temporal identity learning

Collective behavior starts with individual trajectories.

01 / Motivation

To study a school, we need the position and persistent identity of **every individual over time.**

Reliable trajectories enable analysis of spatial organization, local interactions, and abnormal behavior.

2D motion alone cannot recover the full geometry of a 3D school.



Visually homogeneous individuals and mutual occlusion make identity recovery difficult. Motivation image from the paper; original sources credited there: Unsplash / Freepik.

Which detections belong to the same fish?

01 / Problem Formulation

A / Across cameras: spatial association

Match simultaneous 2D observations from different views to reconstruct one 3D position.

B / Across time: temporal association

Link reconstructed positions into one continuous identity, even after short occlusions.

Weak appearance

Many individuals have nearly identical shape and texture.

Frequent overlap

Detections can be missed, noisy, or geometrically ambiguous.

Scarce supervision

Cross-view identities and 3D trajectory labels are costly to annotate.

Solving correspondence in one frame is not enough to preserve identity over a sequence.

Use calibrated geometry as the teacher.

02 / The Key Idea

Learn *who matches whom* from multi-view consistency, then learn *which identity persists* through time.

Required inputs

Synchronized, calibrated cameras and per-frame 2D detections: centroids, boxes, and confidence.

Not required for association training

Cross-view identity labels, temporal identity annotations, appearance descriptors, or ground-truth 3D trajectories.

Self-supervised association and tracking; detector-agnostic, not detector-free.

The missing combination: spatial, temporal, and label-free.

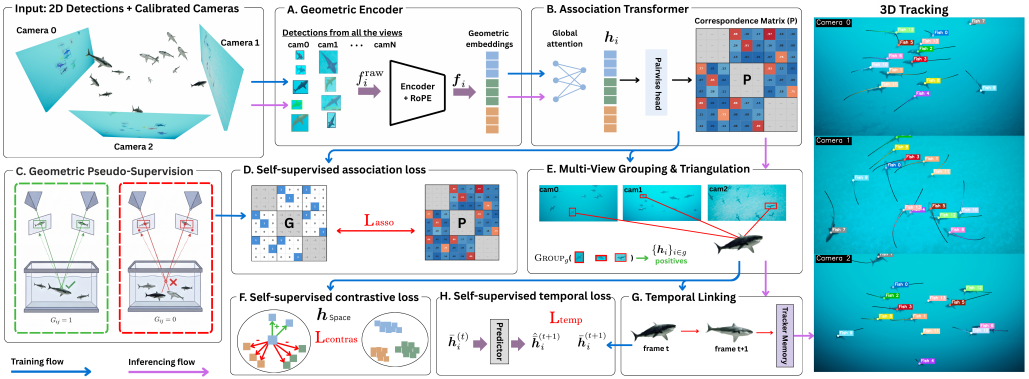
02 / Positioning

Method family	Cross-view	Temporal	Main dependency
Single-view object tracking	No	Yes	Temporal labels / appearance
Supervised multi-view association	Yes	No	Cross-view identities
Self-supervised multi-view association	Yes	No	Visual features
Multi-view object tracking	Yes	Yes	Identity supervision
TrackFish3D	Yes	Yes	Calibrated geometric consistency

Representative families discussed in the paper, rather than an exhaustive statement about every tracker.

From 2D detections to stable 3D identities.

03 / Framework



Paper framework. Geometry supplies correspondence targets; contrastive and temporal objectives shape identity-aware embeddings.

A detection becomes a 9-dimensional geometric token.

03 / Geometric Encoder

$$f_i^{\mathrm{raw}} = [\tilde{u}_i, \tilde{v}_i, \tilde{w}_i, \tilde{h}_i, \tilde{c}_i, s_i, r_{w,x}, r_{w,y}, r_{w,z}]$$

Image geometry

Normalize the mask centroid and box dimensions by image width and height.

View and confidence

Normalize the camera index by $C - 1$ and retain detection confidence.

World-space ray

Back-project the centroid using calibration and normalize the ray direction.

$$\boldsymbol{f}_i = \phi(f_i^{\mathrm{raw}}) \in \mathbb{R}^{128}, \qquad 9 \to 64 \to 128$$

Two-layer MLP + ReLU + LayerNorm.

The association embedding contains no image crop, texture descriptor, or visual re-identification feature.

Resolve a match in the context of the whole scene.

03 / Association Transformer

1 / Collect tokens

All detections from all cameras at one timestamp.

2 / Encode position

Apply 2D rotary position encoding before the encoder stack.

3 / Global attention

Jointly contextualize tokens across the complete frame.

4 / Score pairs

Predict cross-camera same-identity probabilities P_{ij}.

4
Transformer encoder layers

4
Attention heads per layer

128
Token dimensions

Several pairs may be geometrically plausible. Scene-level context helps decide which combinations are consistent.

Triangulate. Reproject. Check consistency.

03 / Geometric Pseudo-Supervision

1 / Candidate pair

Choose two detections in different cameras at the same time.

2 / Triangulate

Estimate the 3D point supported by both observations.

3 / Reproject

Project that point back into the two image planes.

4 / Assign a target

Positive if both reprojections lie inside the corresponding boxes.

$$G_{ij} \in \{0, 1\}, \qquad C_{ij} = \frac{1}{1 + \mathrm{err}_{ij}}$$

Error is averaged across the two views in pixels. Same-camera pairs and self-pairs are masked.

Geometry provides noisy supervision, not guaranteed ground-truth identities.

Learn correspondence without drowning in negative pairs.

03 / Objective 1

$$\mathcal{L}_{\text{asso}} = -\frac{\sum_{G_{ij}=1} C_{ij} \log P_{ij}}{\sum_{G_{ij}=1} C_{ij}} - \frac{\sum_{G_{ij}=0} \log(1 - P_{ij})}{|\{G_{ij} = 0\}|}$$

Positive term

Increase the probability of consistent cross-view pairs. Weight positives by reprojection confidence.

Negative term

Suppress inconsistent pairs. Normalize negatives separately from the positive confidence mass.

Independent normalization prevents abundant negative pairs from dominating the training signal.

Pairwise scores become one validated multi-view hypothesis.

$$A_{ij} = w_p P_{ij} + w_f \frac{\cos(\boldsymbol{h}_i, \boldsymbol{h}_j) + 1}{2}$$

1 / Pairwise matching

Hungarian assignment between camera pairs.

2 / Merge matches

Connect accepted correspondences across views.

3 / Validate in 3D

Triangulate and reject inconsistent reprojections.

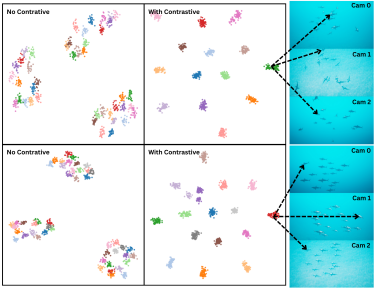
4 / Pool identity

Average member embeddings and L2-normalize.

$$\bar{\boldsymbol{h}}_g = \text{normalize}\left(\frac{1}{|g|}\sum_{i \in g} \boldsymbol{h}_i\right)$$

Same fish, different views: pull together.

03 / Objective 2



Paper visualization: without / with the contrastive objective.

Positives: detections in the same validated multi-view group.

Negatives: detections assigned to other groups in the frame.

$$\mathcal{L}_{\text{ctr}}^{(i)} = -\log \frac{\sum_{j \in \mathcal{P}_i} e^{S_{ij}}}{\sum_{k \neq i} e^{S_{ik}}}$$

S_{ij} is normalized embedding similarity divided by temperature. Average over anchors with positives.

Association scores decide correspondence; contrastive learning makes identity separable in embedding space.

Link identities using both motion and embedding agreement.

03 / Across-Frame Association

$$\mathrm{cost}(i,j) = 0.6\frac{\|\boldsymbol{X}_j^{(t+1)} - \boldsymbol{X}_i^{(t)}\|}{d_{\max}} + 0.4\left(1 - \cos(\bar{\boldsymbol{h}}_i^{(t)}, \bar{\boldsymbol{h}}_j^{(t+1)})\right)$$

1 / Plausible candidates

Use 3D distance and embedding-similarity gates.

2 / Global assignment

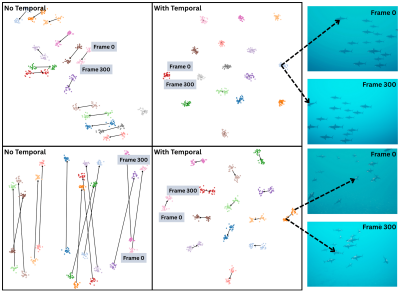
Hungarian matching links groups between adjacent frames.

3 / Reliable matches

Use the minimum confidence of the two matched groups.

Predict the identity embedding, not a visual appearance.

03 / Objective 3



Paper visualization: temporal consistency without / with prediction training.

$$\hat{\bar{\boldsymbol{h}}}_j = \rho(\bar{\boldsymbol{h}}_i^{(t)})$$

$$\mathcal{L}_{\text{temp}} = \frac{\sum_{(i,j)} \boldsymbol{w}_{ij} \| \hat{\bar{\boldsymbol{h}}}_j - \bar{\boldsymbol{h}}_j^{(t+1)} \|^2}{\sum_{(i,j)} \boldsymbol{w}_{ij}}$$

During inference, roll embeddings forward to help reconnect tracks across short gaps.

An identity representation must be consistent across time, not just across cameras.

Learn correspondence first. Add identity consistency next.

03 / Training

$$\mathcal{L} = \mathcal{L}_{\text{asso}} + 0.5\mathcal{L}_{\text{ctr}} + 0.25\mathcal{L}_{\text{temp}}$$

Epochs 1–100

Association warmup.

Learn basic correspondence before forming training groups.

Epochs 101–400

Joint association, contrastive, and temporal learning on consecutive frame pairs.

Grouping provides identity tokens and self-supervised positive sets.

Reported SynFish training configuration from the appendix. No annotated identity or 3D trajectory target enters these objectives.

A trained model directly reconstructs unseen sequences.

03 / Inference

1 / Encode and associate

Frozen network predicts scores and contextual embeddings.

2 / Reconstruct

Group detections and geometrically validate 3D positions.

3 / Track identities

Match observations and propagate missing identity embeddings.

4 / Reconnect

Stitch compatible fragments using identity and 3D proximity.

No test-time optimization, identity annotation, or appearance-based re-identification.

Short-gap interpolation and removal of short-lived ghost tracks are post-processing. They do not introduce new identities.

Dense synthetic schools and real zebrafish.

04 / Evaluation Setup

SynFish / Unreal Engine 5

14 scenes, **4–16 fish**, three synchronized cameras, and seven rig configurations.

1920 × 1080 images.
8 training scenes / 6 held-out test scenes, including one unseen rig.

Synthetic ground-truth 3D trajectories support quantitative evaluation.

3D-ZeF / Real-world zebrafish

Two calibrated cameras; sequences contain **2 or 5 fish**.

2704 × 1520 images.
Train: zebra1 and zebra3.
Test: zebra2 and zebra4, each with 5 fish.

Train self-supervisedly on the real training split; evaluate on separate real sequences.

The real-data result is not a claim of zero-shot synthetic-to-real transfer.

Compare complete 3D tracking pipelines.

04 / Protocol

Four baseline families

- Detections + geometric association, including 3D-SORT.
- Detections + learned association: Self-MVA, ASNet.
- Single-view tracking lifted into 3D: SambaMOTR, MOTIP.
- Multi-camera tracking: ReST, MCTR.

External models are fine-tuned or retrained on the corresponding training splits.

Accuracy and identity stability

MOTA: combines misses, false positives, and identity switches.

MT / **ML:** mostly tracked / mostly lost trajectories.

IDSw / **Frag:** identity switches / trajectory fragmentations.

MTBF: mean time between failures, in frames.

Match reconstructed 3D positions to ground truth under a fixed distance threshold, then evaluate trajectories.

Higher accuracy, with stable identities.

05 / SynFish Results

Selected pipeline	MOTA % ↑	MT % ↑	IDSw ↓	Frag ↓	MTBF ↑
SAM3 + Hungarian	83.5	71.9	7.5	12.8	183.8
SAM3 + 3D-SORT	87.7	77.1	4.0	9.0	209.4
MCTR	83.7	82.7	1.7	14.2	160.6
TrackFish3D / SAM3	95.8	95.8	0.0	6.5	266.3
TrackFish3D / YOLO26+SAM2	96.6	95.8	0.2	3.7	296.5

+8.1 pp
SAM3 MOTA vs. 3D-SORT

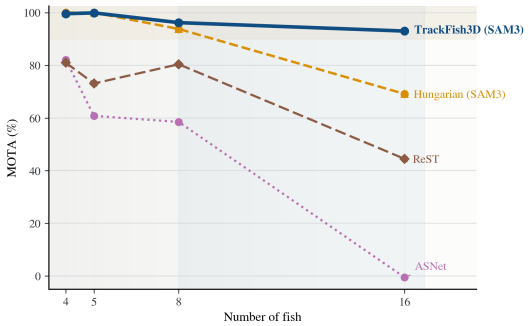
0.0
Mean ID switches / SAM3

95.8%
Mostly tracked / both input variants

Means over the six held-out SynFish scenes. Both detector variants use geometry-only association.

Ambiguity grows with the school.

05 / Density



Paper figure: held-out SynFish scenes with 4–16 fish.

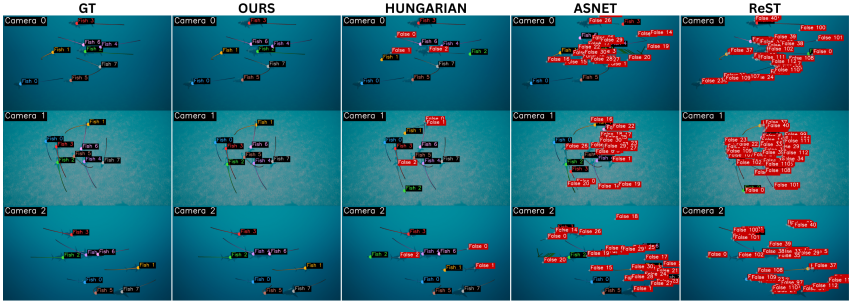
More individuals create more competing cross-view candidates and more overlap.

TrackFish3D remains **above 86% MOTA** on the dense 16-fish test scenes.

Baseline performance deteriorates more strongly at higher density. Fish count is not the only source of difficulty; visibility and rig configuration also vary.

Recover the path without repeatedly changing the identity.

05 / Qualitative Evidence



Columns show ground truth, TrackFish3D, and representative baselines. Colors encode identities; discontinuities reveal fragmentations or identity changes.

Inspect identity continuity alongside the numerical accuracy score.

On real fish, the larger gain is continuity.

05 / Real-World Results

Selected pipeline	MOTA % ↑	IDSw ↓	Frag ↓	MTBF ↑
Hungarian + nearest	77.4	6.0	73.0	42.5
Greedy + nearest	76.9	10.0	76.0	38.7
3D-SORT	75.6	13.5	65.5	43.4
TrackFish3D	81.1	4.0	30.5	93.6

+3.7 pp
MOTA vs. Hungarian + nearest

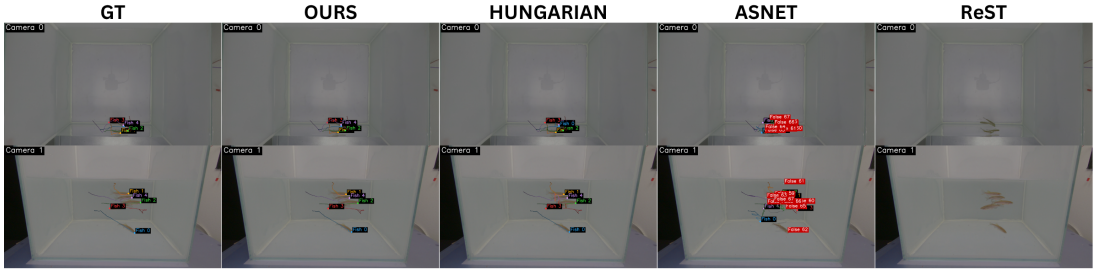
58.2%
Fewer fragmentations

2.20×
Mean time between failures

Means over zebra2 and zebra4. Both test sequences contain five fish.

Temporal identity learning helps through real occlusions.

05 / Real-World Qualitative Evidence



Paper figure: real zebrafish trajectories. TrackFish3D preserves longer, more consistent identity tracks.

ReST produces no reconstructed trajectories in this evaluation because cross-view matching fails. A low switch count alone would not imply successful tracking.

Learn identity from geometry, not appearance.

06 / Takeaways

1 / Geometry supervises

Calibrated multi-view consistency replaces manual identity targets.

2 / Objectives cooperate

Association matches, contrastive learning separates, temporal prediction preserves.

3 / Continuity matters

Accuracy improves alongside identity stability on synthetic and real fish.

Paper	https://arxiv.org/abs/2609.38347
Project, data, videos	https://pattgene.github.io/trackfish3d-projectpage/
Code	https://github.com/pattgene/TrackFish3D
Checkpoints	https://huggingface.co/pattgene/trackfish3d

Thank you!