frontiers in Marine Science

TYPE Original Research
PUBLISHED 28 April 2026
DOI 10.3389/fmars.2026.1749059

Check for updates

OPEN ACCESS

EDITED BY
Ateeq Ur Rehman,
Gachon University, Republic of Korea

REVIEWED BY
Ayodele Periola,
Cape Peninsula University of Technology, South Africa
Hao Wang,
Laoshan National Laboratory, China

*CORRESPONDENCE
H. Seckin Demir
hdemir@asu.edu

RECEIVED 18 November 2025
REVISED 21 March 2026
ACCEPTED 30 March 2026
PUBLISHED 28 April 2026

CITATION
Demir HS, Reavis JL, Blain J, Senko J and Ozev S (2026) Underwater sea turtle behavior recognition: a lightweight pose-to-action pipeline.
Front. Mar. Sci. 13:1749059.
doi: 10.3389/fmars.2026.1749059

COPYRIGHT
© 2026 Demir, Reavis, Blain, Senko and Ozev. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Underwater sea turtle behavior recognition: a lightweight pose-to-action pipeline

H. Seckin Demir[1]*, Janie L. Reavis[2,3], Jennifer Blain[1], Jesse Senko[4] and Sule Ozev[1]

[1]Department of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, United States, [2]School of Life Sciences, Arizona State University, Tempe, AZ, United States, [3]Inwater Research Group, Jensen Beach, FL, United States, [4]School for the Future of Innovation in Society, Arizona State University, Tempe, AZ, United States

Automated recognition of animal behaviors is an important computer vision task that improves ecological monitoring and behavioral analysis. Compared to generic human action recognition, these applications often suffer from severe constraints such as low onsite computational power, limited data availability for training learning-based models, and suboptimal image quality due to environmental conditions. For sea turtles, behavior in relation to fishing gear is particularly important for understanding and reducing the bycatch, or incidental take, and associated mortality. Monitoring such behavior underwater is challenging because viewing angles vary over time, and pose and motion trajectories are highly dependent on the camera angle. In this study, we address this problem with a compact pose-to-action pipeline that detects a small set of turtle morphological keypoints in each frame, and then classifies short sequences of keypoints into U-turn, reversal, or other routine behaviors. While we employ the YoloV8n pose model for keypoint detection, we use a shallow fully connected network for classifying the behavior types. Our two-stage training strategy allows us to train our pose estimation network with real data while optimizing the behavior recognition network with both real annotated clips and a large set of simulated trajectories including various camera geometries and motion parameters. We further reduced the computational requirements by finding a balance between the input frame rate and recognition accuracy. Our experimental results show that we can achieve 93.2% recognition accuracy with a minimum frame rate requirement of 2.86fps.

KEYWORDS
bycatch reduction, low-power inference, pose estimation, sea turtle behavior, underwater

1 Introduction

Video-based action recognition has advanced rapidly with deep learning. Modern 2D and 3D convolutional networks, as well as more recent transformer-style architectures, can model complex spatiotemporal patterns in human activities and achieve strong performance on large benchmarks such as Kinetics (Carreira and Zisserman, 2017; Feichtenhofer et al., 2019). However, these models are typically trained on massive labeled datasets, operate directly on dense pixel inputs, and assume server-scale compute. Many animal behavior applications do not meet these assumptions: they must function with limited data, domain-specific viewpoints, and strict computational budgets.

Recognition of animal behaviors is important for precision livestock farming, laboratory ethology, and wildlife conservation. A recent study has used video sequences and deep models to classify sow estrus behaviors (Lei et al., 2025). Another study uses a dual-view multi-scale spatiotemporal fusion network and a lightweight 3D fitting stage to quantify macaque pose and actions (Chen et al., 2026). Fazzari et al. surveys different pipelines for animal behavior analysis methods and notes that data limits and hardware limits often decide which route works in practice (Fazzari et al., 2025). These findings push us toward a pose-to-action design that stays light enough for our proposed task: sea turtle behavioral monitoring.

Monitoring the behavior of sea turtles in aquatic facilities and natural habitats is a significant part of conservation practice. From a conservation perspective, bycatch, or incidental take, of sea turtles in coastal gillnet fisheries remains a major concern, and understanding the fine-scale behaviors in relation to fishing gear is a critical step toward designing and improving effective bycatch reduction technologies (Southwood et al., 2008). Our previous top-view study demonstrated that a compact set of spatiotemporal motion features derived from visual tracking and orientation estimates can reliably identify U-turn and reversal behaviors linked to gear interactions (Reavis et al., 2021). That system benefited from a camera mounted above the test tank, a largely planar motion pattern, and carefully controlled lighting.

Underwater observation is a more challenging task due to several reasons. The motion unfolds in three dimensions. With turtles simultaneously changing depth, roll, and pitch, the same maneuver can look dramatically different depending on camera placement. The apparent trajectory is also tightly coupled to camera pose. Moreover, factors such as refraction, turbidity, color shift, and specular glare can degrade image quality in underwater footage. These factors complicate both reliable pose estimation and behavior recognition, particularly when an ideal system is expected to operate continuously on low-power hardware deployed at remote field sites or attached to autonomous recording units.

Recent image-analysis studies in challenging visual domains also point to the value of richer feature fusion and structured representations. In underwater enhancement, WaterCycleDiffusion combines visual and textual cues to improve restoration under severe color distortion and paired-data scarcity (Wang et al., 2026). In adjacent sensing domains, S2G-GCN shows that graph-based modeling can compactly encode structured signal relationships (Li et al., 2025), while a visual-textual mutual guidance fusion network for remote sensing visual question answering demonstrates how cross-modal interaction can strengthen semantic reasoning in cluttered scenes (Lu et al., 2026). These studies are not direct behavior-recognition baselines, but they reinforce a broader design principle that is relevant here: when imaging conditions are difficult and semantics are subtle, carefully chosen representations and fusion mechanisms matter as much as raw detector capacity.

To address these constraints, we adopt a pose-to-action strategy in which a lightweight keypoint detector provides a compact representation of turtle pose and a small temporal network maps sequences of keypoints to behavior labels. A YOLOv8n-pose model (Jocher et al., 2023) predicts four anatomical keypoints which are head, tail, and the left and right body sides. These keypoints are chosen to capture heading, curvature, and body orientation with minimal annotation cost. A compact fully connected network processes short temporal sequences of these keypoints to classify U-turn, reversal, and other routine behaviors. To determine the minimal temporal context that preserves accuracy, we combine an annotated underwater dataset with a large simulation dataset that systematically varies camera geometry and motion parameters.

This paper makes three contributions. First, we extend prior top-view work on turtle behavior recognition to the underwater setting, introducing a four-keypoint representation and temporal classifier that handle viewpoint changes while remaining suitable for embedded deployment. Second, we develop a simple but effective simulation framework that generates synthetic keypoint trajectories under randomized camera poses and motion parameters, allowing us to analyze the trade-off between temporal window length and recognition accuracy. Third, we demonstrate that ten frames of context are sufficient to reach high accuracy on U-turn and reversal behaviors, corresponding to a very low effective sampling rate that can be achieved on an edge device such as Raspberry Pi. Figure 1 summarizes the overall pipeline.

1.1 Related work

Our design builds on three areas: markerless animal pose estimation, mapping pose to behavior with temporal models, and underwater visual sensing. Markerless pose estimation has become the de facto foundation for behavior analysis across species. DeepLabCut demonstrated that transfer learning with deep networks enables accurate pose estimation from modestly sized, user-defined datasets (Mathis et al., 2018). Subsequent systems such as DeepPoseKit (Graving et al., 2019) and SLEAP (Pereira et al., 2020) improved multi-animal tracking and throughput, while idtracker.ai addressed scalable identity tracking in groups without markers (Romero-Ferrero et al., 2019). These tools established that compact, semantically meaningful keypoints can support downstream analyses with limited annotation budgets.

Translating pose into behavior labels spans supervised and unsupervised paradigms. Early interactive pipelines such as JAABA (Kabra et al., 2013) empowered domain experts to curate features and train behavior classifiers. Unsupervised approaches, for example motion sequencing in MoSeq (Wiltschko et al., 2015), discover stereotyped sub-second motifs that compose behavior. More recently, end-to-end supervised video models like DeepEthogram (Bohnslav et al., 2021) leverage deep features and temporal context from pixels directly. Complementary to these appearance-heavy models, skeleton-based action recognition has shown that operating on keypoints alone can be both discriminative and efficient. Graph-based methods such as ST-GCN (Yan et al., 2018) and 2s-AGCN (Shi et al., 2019) capture spatial relations and temporal dynamics over joints with strong performance and reduced compute compared with 3D CNNs. Efficient temporal operators like the Temporal Shift Module further reduce the cost of video modeling, informing design choices for resource-limited settings (Lin et al., 2019).

Underwater imaging introduces domain-specific challenges that are not present in aerial or top-view monitoring. Refraction,

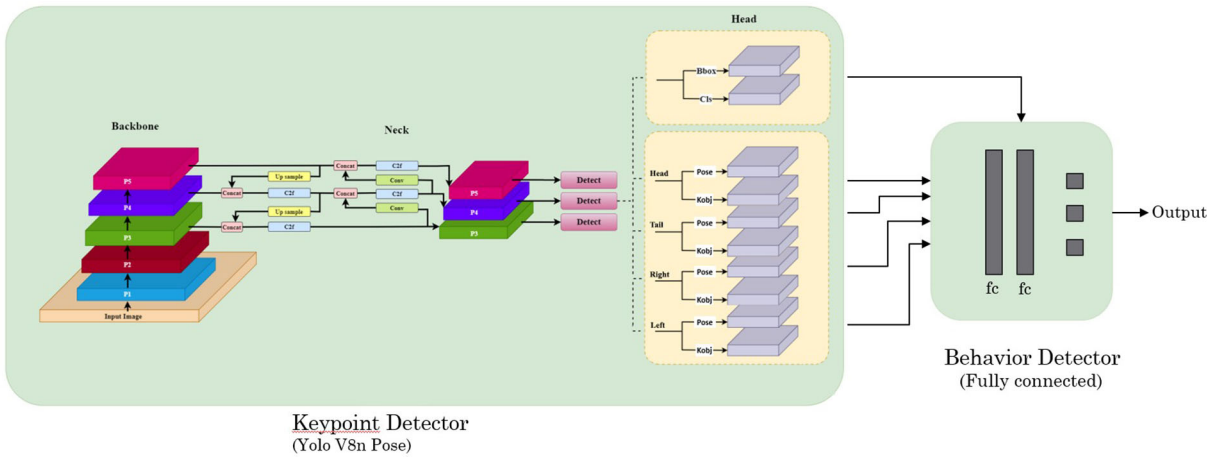

FIGURE 1
Pipeline overview. Frame-level keypoints from YOLOv8n-pose are buffered into windows of *N* frames and classified by a lightweight temporal network. The resulting behavior labels can be aligned with experimental events or gear configurations to study turtle responses under different conditions.

turbidity, color shift, and low contrast can degrade both detection and keypoint stability. Surveys on underwater image enhancement document these degradations and restoration strategies (Anwar and Li, 2020; Demir et al., 2024). Large-scale efforts such as Fish4Knowledge (Fisher et al., 2016) have curated long-term coral reef footage and analytical tools, which emphasizes the variability of underwater scenes and the need for robustness. While there is emerging work on underwater object detection and species monitoring using deep learning, including sea turtle detection in the wild (Putro et al., 2024), there is comparatively little on underwater pose-to-behavior pipelines tailored for embedded, continuous operation.

Among recent spatiotemporal detectors, YOLO-Act unifies action localization with multi-frame processing, highlighting how one-stage detectors can be adapted to temporal reasoning (Alzahrani et al., 2025). In precision livestock monitoring, spatio-temporal modeling for sow estrus shows that compact motion descriptors and temporal classifiers can reach high accuracy without heavy 3D CNNs (Lei et al., 2025). These directions inform our emphasis on a minimal keypoint representation and a small temporal head that preserves accuracy while remaining deployable.

Recent behavior pipelines usually pick between a pose-to-action route and an appearance-heavy video model. Fazzari et al (Fazzari et al., 2025). and Arablouei et al (Arablouei et al., 2023a, b) point out how these choices affect accuracy, compute, and sensor fusion. Pose models trade detail for speed, while appearance models keep pixels to gain robustness. Our method follows the pose route to meet edge limits, and we rely on simulation to keep the temporal cues that heavier video models would otherwise cover.

Relative to our previous top-view study (Reavis et al., 2021), the present work targets the added variability of underwater viewpoints while retaining a minimal pose representation. Compared with appearance-driven approaches in the broader action-recognition and animal-behavior literature (Carreira and Zisserman, 2017; Feichtenhofer et al., 2019; Alzahrani et al., 2025; Lei et al., 2025; Chen et al., 2026), our approach emphasizes a compact four-keypoint skeleton and a short temporal window. This choice allows us to trade heavy spatiotemporal computation for a lightweight classifier over keypoints, without sacrificing the pose cues most predictive of U-turns and reversals. On the detection side, we adopt a YOLOv8 pose head (Jocher et al., 2023) for per-frame keypoints and shift most of the temporal burden to the behavior network. This decoupling is central to meeting throughput constraints on Raspberry Pi class devices while preserving accuracy.

2 Materials and methods

We designed the study to achieve a balance between behavioral coverage, robustness to underwater variability, and feasibility on low-power hardware. The pipeline contains per-frame pose estimation, temporal aggregation of keypoints, and behavior classification over short clips (Figure 1). Our training strategy includes two parts. To train the keypoint detector, we used real underwater sea turtle images with manually annotated keypoints. In the second step, we used underwater video clips captured at our test facilities, where each of these clips corresponds to a behavior class. For each video clip, we extracted the keypoints for all the frames and formed a temporal sequence of keypoints. To further extend the behavior dataset, we used a large set of simulated keypoint trajectories.

2.1 Keypoint detection dataset

To support pose estimation, we manually annotated 16,000 underwater frames with a bounding box and four anatomical keypoints: head, tail, and the left and right body sides. The dataset contains samples from both the tank setup and other randomly selected sea turtle images to prevent over-fitting. Annotators followed guidelines that fixed a consistent left-right convention in camera coordinates and prescribed how to mark partially occluded or low-contrast keypoints. Ambiguous frames were revisited, and spot checks were used for quality control. These annotations were used to train a YOLOv8n-pose model specialized to this four-keypoint structure. Figure 2 illustrates sample annotations in the dataset.

2.2 Real-world dataset for behavior recognition

Underwater videos were recorded in a test tank using a GoPro Hero 8 positioned below the waterline at varying heights and angles to induce viewpoint changes. Subjects were green (*Chelonia mydas*) and loggerhead (*Caretta caretta*) sea turtles ranging from 25cm to 78cm straight carapace length. These subjects were participating in behavioral binary choice experiments examining sea turtle interactions with fishing gear and bycatch reduction technologies. Lighting and water clarity varied naturally across recording sessions, resulting in frames with reduced contrast and color shifts. We focused on three behavior classes, two of which have been identified in previous studies: U-turn, reversal, and other (Reavis et al., 2021, 2025). U-turns involve a sustained change in heading and body curvature, which results in a smooth arc. Reversals are characterized by a backwards motion typically observed after interacting with the test net. The class "other" includes straight or gently curved swimming and idle motion without decisive heading change. We prepared 40 real sequences per class to have a balanced dataset. A sample subset of the underwater video captures is publicly available for reference [1].

To ensure a consistent protocol with our earlier top-view system (Reavis et al., 2021), we followed the same animal handling, recording schedule, and behavior definitions. Recordings were performed at the same facility, using the same tank, cameras matched in resolution and frame rate, and identical lighting policy. The stimuli and environmental conditions used to elicit and observe routine swimming behaviors were similar to the prior setup. A sample underwater image and the top view of the test tank are shown in Figure 3.

The overall dataset and splits are summarized in Table 1.

2.3 Simulation framework

To systematically analyze the effect of temporal context and to obtain a more robust model against different viewing angles, we developed a simple kinematic simulator that generates keypoint trajectories for the three behavior classes. For each class, 10,000 simulations randomized camera angle and position, field of view, and motion parameters such as speed, turn diameter, and the addition of small perturbation in all axes. The simulator outputs sequences of four keypoints projected into image coordinates, and adds a small positional shift to model localization error. These synthetic sequences were used to study the accuracy-latency trade-off as a function of the number of frames N and to add diversity to training. Figure 4 illustrates the keypoint trajectories for a sample u-turn simulation and their image-plane projections.

2.4 Augmentation and dataset assembly

We expanded the simulation dataset using augmentation techniques to increase robustness to underwater variability. Horizontal flips were applied with appropriate remapping of left and right keypoints. Additional transforms included random noise injection and multi-scale cropping, which are intended to help the detector and temporal model generalize across distance and partial views. After augmentation, the training set comprised approximately 180,000 samples. We adopted a 70/20/10 split at the sequence level to avoid leakage, and used the test portion for all final metrics.

2.5 Pose estimation

We trained a YOLOv8n-pose model to predict the four keypoints per frame from underwater images. The nano variant provides a balance between accuracy and efficiency for embedded use. Architecturally, YOLOv8 combines a lightweight C2f backbone with a Path Aggregation Feature Pyramid Network (PAN– FPN) neck that aggregates multi-scale features, and an anchor-free, decoupled head. The pose variant attaches a keypoint regression branch to the detection head so that, for each detected turtle, the network outputs a bounding box, objectness score, and the image coordinates of K keypoints; here $K = 4$ (head, tail, left, right) (Jocher et al., 2023). We normalize keypoint targets to the input size to improve transfer across scales. Training uses the annotated frames and the same augmentations described above; the loss combines objectness and box regression with a visibility-weighted regression over keypoints. At inference, per-frame keypoints are mapped back to pixel coordinates and lightly smoothed to reduce jitter before temporal aggregation.

Although newer YOLO-family variants have continued to appear, our choice of YOLOv8n-pose was driven by the specific constraints of this study rather than by a claim that it is universally the strongest detector. The primary goal here is a deployable underwater pose-to-action pipeline for embedded inference, so we prioritized a model with an established native pose head, stable training and export tooling, and a compact nano-scale configuration that can be integrated with a four-keypoint custom dataset at low latency. In this context, YOLOv8n-pose provided a practical accuracy-efficiency trade-off and a reproducible implementation path for edge-oriented deployment. A broader comparison against newer YOLO variants under the same four-keypoint and embedded-compute constraints is a useful direction for future work.

Training setup and parameters are provided in Table 2.

2.6 Temporal subsampling strategy

Instead of processing all the images at full frame rate of the camera (typically 30fps), we are aiming to find the minimum frame rate that can still give a reasonable behavior recognition accuracy. Using this approach, we can reduce the required processing power and enable the method to run on edge devices.

To investigate the relation between the number of used frames and the corresponding recognition accuracy, we use a subsampling strategy. In this approach, we use N frames throughout the whole behavior sequence as illustrated in Figure 5. The purpose is to find the minimum value of N that gives a successful behavior recognition result over the whole dataset.

2.7 Behavior network

The classifier operates directly on the temporal stack of keypoints as an $N\times4\times2$ tensor of image coordinates. We used a

1 https://osf.io/62v3c/files/osfstorage?view_only=8b771072a98d4feeb8d70f928f4c6b55

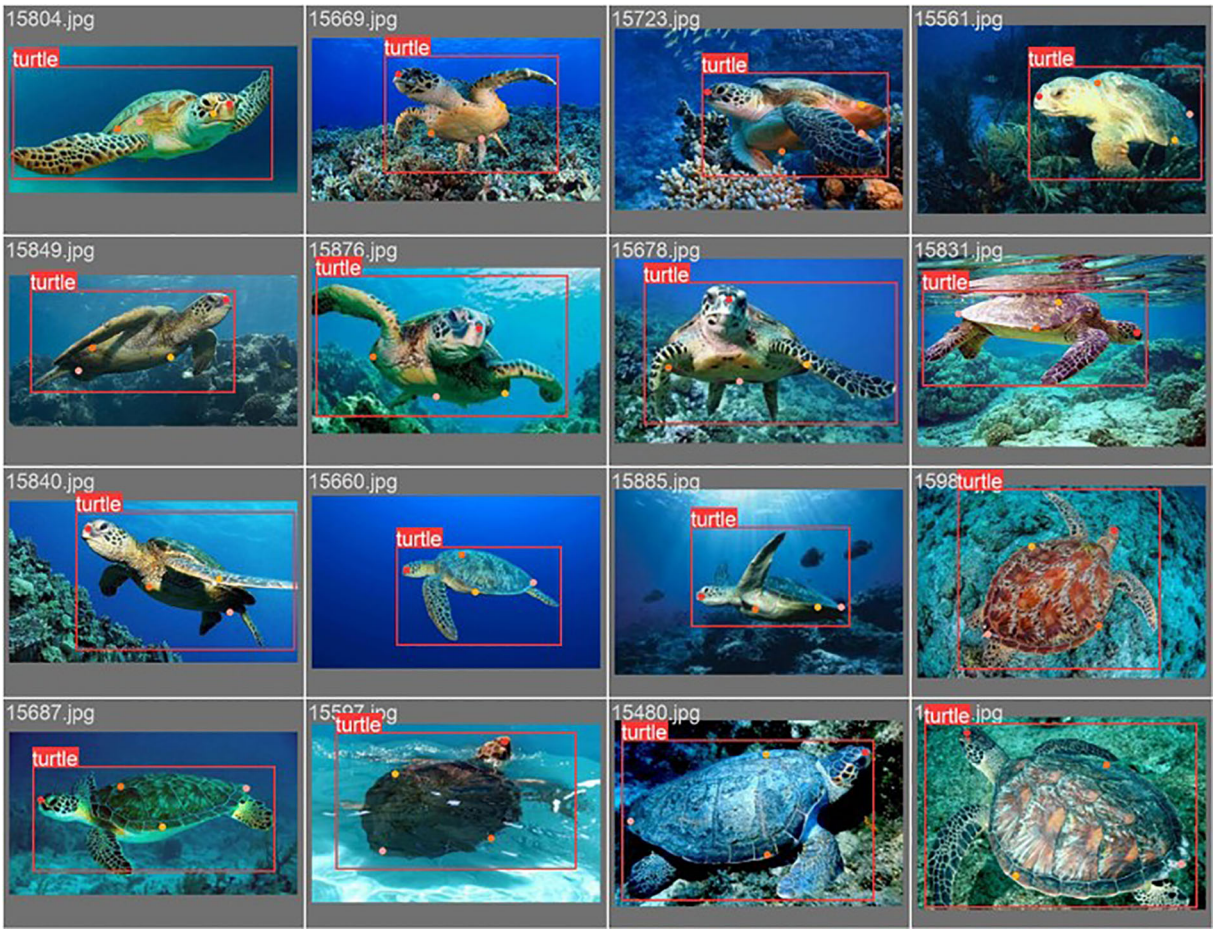

FIGURE 2
Data and annotation examples. Underwater frames with bounding boxes and four keypoints (head, tail, left, right). Examples span viewpoint changes, moderate turbidity, and partial occlusions.

shallow fully connected neural network with two layers to transform this sequence into class probabilities. The network begins with a flattening over time and forms a low-dimensional descriptor per clip, followed by rectified linear units that encourage sparse, interpretable responses to curvature and heading-change patterns. This design keeps parameters small and latency low while retaining the capacity to discriminate temporal features of different behavior types. Missing or low-confidence keypoints are handled by light smoothing and, when necessary, short linear interpolation. Training employed the same 70/20/10 split, with early stopping on validation accuracy.

2.8 Training and implementation details

We split data at the sequence level into 70%/20%/10% for training, validation, and testing to avoid temporal leakage across sets. Hyperparameters were tuned on the validation set, and the final model selection used early stopping on validation accuracy.

FIGURE 3
A sample underwater image and corresponding top view of the experimental setup. (The same tank, camera placement, lighting policy, and recording schedule were used for both the previous top-view study and the underwater recordings reported here).

TABLE 1 Behavior dataset summary that shows available real sequences and final augmented samples.

Real sequences	U-turn 40	Reversal 40	Other 40
Minimum video length(s)	3.8	3.5	3.65
Maximum clip length (s)	17.7	14.6	15.0
Synthetic sequences	10K	10K	10K
Total (inc. augmentation)	60K	60K	60K
Train/Val/Test split (%)		70/20/10	

We adopted standard optimizers and learning-rate schedules suitable for small models, and we matched data augmentations between pose and behavior training to maintain consistency. During inference, frames were processed by the pose model and keypoints were buffered into sliding windows of length N. Per-window predictions were smoothed to reduce flicker at behavior boundaries. All results reported in this paper reflect the held-out test split.

3 Results

3.1 Single-frame detection

The YOLOv8n-pose model produced stable keypoint predictions across a wide range of underwater conditions, including moderate turbidity, changing lighting, minor glare, and variation in viewing angle. Head and tail locations were generally consistent, while left–right estimates occasionally drifted during rapid rolls or strong specular highlights. Light temporal smoothing corrected most of these cases.

Quantitatively, the model achieved a precision of 0.988, recall of 0.992, mAP50 of 0.985, and mAP50–95 of 0.967 on the held-out test set (Table 3). These results indicate that a compact four-keypoint representation can be extracted reliably from single underwater frames and can serve as a stable input for behavior classification in the next step.

3.2 Temporal subsampling and behavior recognition

To evaluate how much temporal context is needed for accurate behavior recognition, we varied the number of frames N used per sequence for classification. Using the combined real and simulated dataset, accuracy increased steadily with larger N. As the Table 4 summarizes, the classifier reaches 39.3% accuracy with only 3 frames. This improved to 55.4% at $N = 5$ and 84.1% at $N = 8$. Performance peaked at $N = 10$, where accuracy reached 93.2%. Increasing the window further to 12 frames did not yield additional improvements, suggesting that 10 frames provide sufficient information for distinguishing U-turns, reversals, and other movements.

Figure 6 shows that the confusion matrix for $N = 10$ provides strong class separation. Most errors occurred during brief transitional movements where a turtle made a small arc or a partial change in heading that resembled the early phase of a reversal or a U-turn. These cases often appeared ambiguous even during manual

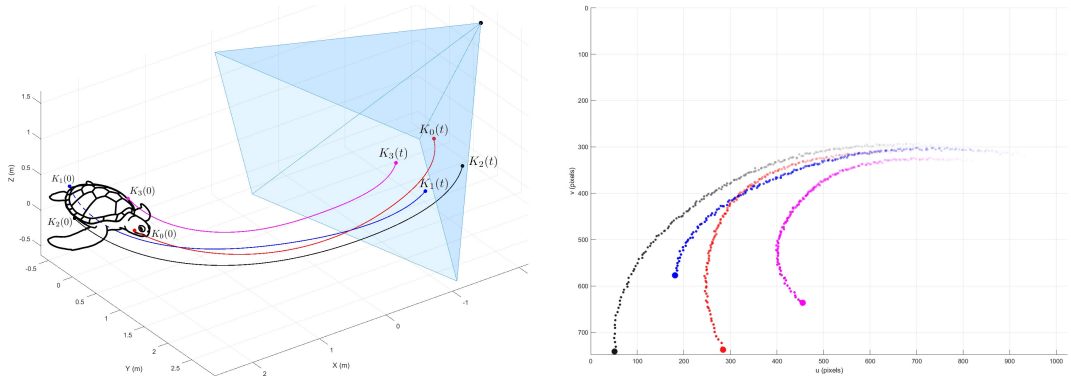

FIGURE 4
A sample from the simulated behavior dataset. Left: 3D trajectory of keypoints for a u-turn simulation. Right: image-plane projection under a randomized camera pose; the four keypoints provide heading and curvature cues.

TABLE 2 YOLOv8n-pose training setup.

Input resolution	640 × 640
Optimizer/LR schedule	Adam, cosine decay
Batch size/epochs	32/100
Augmentations	flips, noise, multi-scale crops
Losses	objectness, box regression, keypoint regression

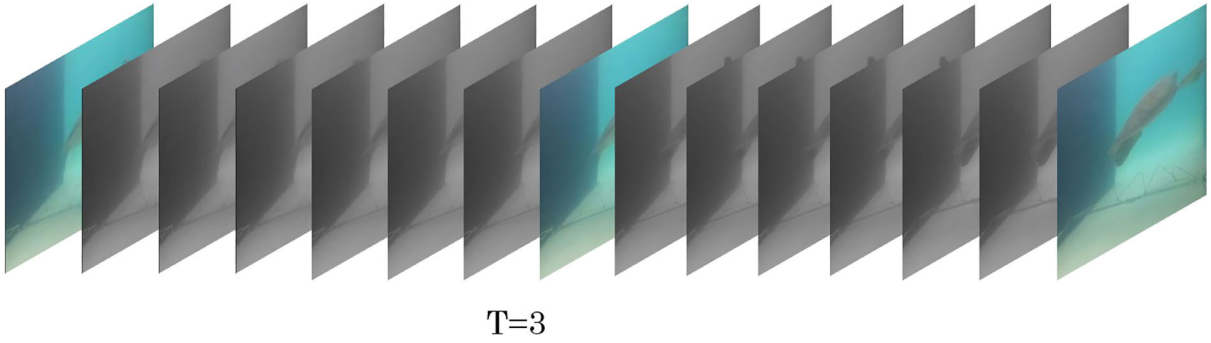

FIGURE 5
Illustration of the temporal subsampling strategy. Only three pose frames are used across the full sequence which enables reduction in computational need.

review, indicating that some misclassifications reflect the subtlety of

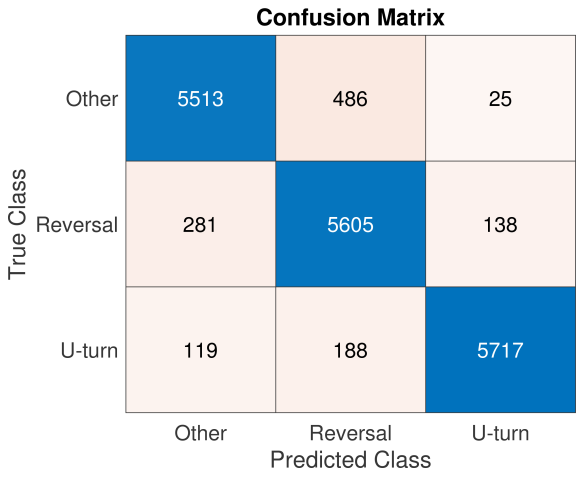

FIGURE 6
Confusion matrix for the three classes for $N = 10$.

the behaviors rather than systematic model bias.

3.3 Runtime considerations and edge deployment

The short temporal window required for accurate behavior classification significantly reduces computational load. Considering a 5 seconds clip at 30fps, this would correspond to processing 10 frames instead of 150 frames to recognize the behavior. Considering the worst case (shortest sequence) in our dataset, $N = 10$ translates to a temporal sampling requirement of only 2.86 fps for the classifier. This requirement is far below the processing capacity of modern lightweight edge devices. For example, the YOLOv8n model runs at about 20 fps on a Raspberry Pi 5 [2]. As the fully connected classifier adds negligible latency, the overall system can operate continuously with low computational power demand.

TABLE 3 **Single frame sea turtle detection results.**

Precision	Recall	mAP50	mAP50-95
0.988	0.992	0.985	0.967

TABLE 4 **Accuracy versus temporal length *N* for behavior classification.**

N	3	5	8	10	12
Accuracy (%)	39.3	55.4	84.1	93.2	92.8

4 Discussion

The results of this study show that a compact pose-based action recognition pipeline can classify underwater sea turtle maneuvers with high accuracy, even when video is sampled at relatively low frame rates. This finding has important implications for marine behavioral ecology and conservation, as the referred U-turn and reversal behaviors have been noted as indicators of how turtles

2 https://github.com/Qengineering/YoloV8-ncnn-Raspberry-Pi-5

interact with fishing gear (Reavis et al., 2025). Automated identification of these maneuvers enables researchers to quantify such movements at scales that would be very labor-intensive to annotate manually. With continuous monitoring, it becomes possible to characterize the sequence of behaviors that precede and follow gear interactions, to compare behavioral patterns across gear types or experimental treatments, and to study how animals respond to stimuli designed to reduce bycatch risk.

The practical value of low-frame-rate classification is especially clear when considering deployment on gear-mounted cameras. Devices placed directly on nets or other fishing gear must operate under strict energy and memory constraints, and high-resolution, high-speed video is rarely feasible. Demonstrating that accurate behavior recognition is possible with only a small number of frames suggests that low-power edge devices can be deployed for underwater behavioral monitoring.

A notable strength of this work lies in the dual-dataset strategy combining real underwater behavior with simulated trajectories. Real footage provides irreplaceable biological authenticity, while the simulations allow us to explore conditions that are difficult to reproduce experimentally, such as extreme viewing angles, variable speed profiles, or systematic perturbations in pose estimation. This combination enables a more rigorous assessment of how temporal subsampling affects classification performance, and provides a clearer picture of the system's robustness under a wider set of scenarios. Such an approach may be useful for other marine species as well, where collecting large, labeled behavioral datasets is a significant challenge.

Several limitations should be acknowledged. The real behavioral sequences used in this study were recorded in a controlled tank environment, and natural conditions in the wild will inevitably introduce additional complexity, including occlusions, turbidity, rapidly changing backgrounds, and interactions with other species. Moreover, the behavior classes we consider represent only a small set; future models may incorporate more fine-grained behaviors, including approach patterns, escape swimming, entanglement attempts, or collision dynamics. Continued work is also needed to collect larger annotated datasets and to validate the system on active fishing gear in the field.

Despite these limitations, this study provides an important proof of concept by demonstrating that biologically meaningful behaviors can be recognized from underwater footage using a simple, efficient model tailored to the constraints of conservation monitoring. The ability to quantify U-turn and reversal behaviors helps researchers study how sea turtles interact with fishing gear and evaluate the effectiveness of mitigation strategies. As these tools are adapted and extended in future fieldwork, they have the potential to strengthen both the scientific understanding of sea turtle behavior and the practical efforts to reduce bycatch in coastal fisheries.

Data availability statement

The data presented in the study are deposited in the repository: "https://osf.io/62v3c/files/osfstorage?view_only=8b771072a98d4feeb8d70f928f4c6b55".

Ethics statement

The animal study was approved by The Florida Fish and Wildlife Conservation Commission. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

HD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. JR: Data curation, Resources, Writing – review & editing. JB: Funding acquisition, Project administration, Supervision, Writing – review & editing. JS: Funding acquisition, Project administration, Supervision, Writing – review & editing. SO: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This material is based upon work supported by the National Science Foundation under Grant No. 1837473. The research is also supported by Inwater Research Group and Florida Power and Light Company. Florida Power and Light Company was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Generative AI was used as a writing assistant to help improve the clarity, flow, and organization of the manuscript text. It supported tasks such as rephrasing sentences for readability, checking grammar, suggesting clearer wording, and helping restructure some sections while keeping all scientific content, data, and interpretations provided by the authors. No parts of the scientific results, analyses, figures, or conclusions were generated by AI. All research design, data collection, model development, experiments, and findings are entirely the work of the authors.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alzahrani, N., Bchir, O., and Ismail, M. M. B. (2025). Yolo-act: Unified spatiotemporal detection of human actions across multi-frame sequences. *Sensors* 25, 3013. doi: 10.3390/s25103013

Anwar, S., and Li, C. (2020). Diving deeper into underwater image enhancement: A survey. *Signal. Processing: Image Communication* 89, 115978. doi: 10.1016/j.image.2020.115978

Arablouei, R., Moeller, L., Kusy, B., Kottege, N., and Gordon, N. (2023a). Animal behavior classification via deep learning on embedded systems. *Comput. Electron. Agric.* 207, 107707. doi: 10.1016/j.compag.2023.107707

Arablouei, R., Moeller, L., Kusy, B., Kottege, N., and Gordon, N. (2023b). Multimodal sensor data fusion for in-situ classification of animal behavior using accelerometry and gnss data. *Smart Agric. Technol.* 4, 100163. doi: 10.1016/j.atech.2022.100163

Bohnslav, J. P., Wimalasena, N. K., Clausing, K. J., Dai, Y. Y., Yarmolinsky, D. A., Cruz, T., et al. (2021). Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *elife* 10, e63377. doi: 10.7554/elife.63377

Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, Z., Jiang, Z., Guan, Q., and Ma, X. (2026). End-to-end motion detection via multi-scale spatialtemporal feature fusion for dual-view 3d macaque behavior quantification. *Expert Syst. Appl.* 299, 129765. doi: 10.1016/j.eswa.2025.129765

Demir, H. S., Blain Christen, J., and Ozev, S. (2024). "Underwater image restoration using light attenuation", in *Advances in Visual Computing: 19th International Symposium, ISVC 2024, Lake Tahoe, NV, USA, October 21–23, 2024, Proceedings, Part II* (Berlin, Heidelberg: Springer-Verlag), 281–291. doi: 10.1007/978-3-031-77389-1_22

Fazzari, E., Romano, D., Falchi, F., and Stefanini, C. (2025). Animal behavior analysis methods using deep learning: A survey. *Expert Syst. Appl.* 289, 128330. doi: 10.1016/j.eswa.2025.128330

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). "Slowfast networks for video recognition", in *Proceedings of the IEEE International Conference on Computer Vision*.

Fisher, R. B., Chen-Burger, Y. H., Giordano, D., Hardman, L., and Lin, F. P. (2016). "Fish4knowledge: Collecting and analyzing massive coral reef fish video data," in *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*. Vol. 104. (Berlin/Heidelberg, Germany: Springer). doi: 10.1007/978-3-319-30208-9

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., et al. (2019). Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife* 8, e47994. doi: 10.7554/elife.47994

Jocher, G., Chaurasia, A., and Qiu, J. (2023). *Ultralytics yolov8*. Available online at: https://github.com/ultralytics/ultralytics.

Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S., and Branson, K. (2013). Jaaba: Interactive machine learning for automatic annotation of animal behavior. *Nat. Methods* 10, 64–67. doi: 10.1038/nmeth.2281

Lei, K., Li, B., Yang, H., Wang, H., Wang, D., and Xiong, B. (2025). Spatiotemporal modeling and intelligent recognition of sow estrus behavior for precision livestock farming. *Animals* 15, 2868. doi: 10.3390/ani15192868

Li, X., Sun, W., Ji, Y.-G., and Huang, W. (2025). S2g-gcn: A plot classification network integrating spectrum-to-graph modeling and graph convolutional network for compact hfswr. *IEEE Geosci. Remote Sens. Lett.* 22, 1–5. doi: 10.1109/LGRS.2025.3623931

Lin, J., Gan, C., and Han, S. (2019). "Tsm: Temporal shift module for efficient video understanding", in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE), 7082–7092. doi: 10.1109/iccv.2019.00718

Lu, X., Wang, H., Bai, L., Wang, M., and Ren, P. (2026). A visual-textual mutual guidance fusion network for remote sensing visual question answering. *Pattern Recognit.* 176, 113258. doi: 10.1016/j.patcog.2026.113258

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., et al. (2018). Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* 21, 1281–1289. doi: 10.1038/s41593-018-0209-y

Pereira, T. D., Tabris, N., Li, J., Ravindranath, S., Papadoyannis, E. S., Wang, Z. Y., et al. (2020). Sleap: Multi-animal pose tracking. *BioRxiv* 2020–0. doi: 10.1101/2020.08.31.276246

Putro, M. D., Mose, Y., Andaria, A. C., Litouw, J., Poekoel, V. C., and Najoan, X. (2024). Streamlining deep learning network for real-time sea turtle detection. *Jurnal Rekayasa Elektrika* 20, 116–124. doi: 10.17529/jre.v20i3.35236

Reavis, J. L., Demir, H. S., Witherington, B. E., Bresette, M. J., Blain Christen, J., Senko, J. F., et al. (2021). Revealing sea turtle behavior in relation to fishing gear using color-coded spatiotemporal motion patterns with deep neural networks. *Front. Mar. Sci.* 8. doi: 10.3389/fmars.2021.785357

Reavis, J. L., Senko, J. F., Demir, H. S., Witherington, B. E., Blain Christen, J., Bresette, M. J., et al. (2025). Novel behavioral responses of sea turtles to gillnet fishing gear. *Biol. Conserv.* 306, 111161. doi: 10.1016/j.biocon.2025.111161

Romero-Ferrero, F., Bergomi, M. G., Hinz, R. C., Heras, F. J. H., and de Polavieja, G. G. (2019). idtracker.ai: Tracking all individuals in small or large collectives of unmarked animals. *Nat. Methods* 16, 179–182. doi: 10.1038/s41592-018-0295-5

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). "Two-stream adaptive graph convolutional networks for skeleton-based action recognition", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE), 12018–12027. doi: 10.1109/cvpr.2019.01230

Southwood, A., Fritsches, K., Brill, R., and Swimmer, Y. (2008). Sound, chemical, and light detection in sea turtles and pelagic fishes: Sensory-based approaches to bycatch reduction in longline fisheries. *Endanger. Species Res.* 5, 225–238. doi: 10.3354/esr00097

Wang, H., Zhang, W., Xu, Y., Li, H., and Ren, P. (2026). Watercyclediffusion: Visual-textual fusion empowered underwater image enhancement. *Inf. Fusion* 127, 103693. doi: 10.1016/j.inffus.2025.103693

Wiltschko, A. B., Johnson, M. J., Iurilli, G., Peterson, R. E., Katon, J. M., Pashkovski, S. L., et al. (2015). Mapping sub-second structure in mouse behavior. *Neuron* 88, 1121–1135. doi: 10.1016/j.neuron.2015.11.031

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proc. AAAI Conf. Artif. Intell.* 32. doi: 10.1609/aaai.v32i1.12328