ISSN: 2349-6002

Deep Learning-Based Underwater Object Detection and Tracking with An Intelligent Chatbot Interface

Dr. A. Jagan[1], Arunkumar S[2], Ragunath D[3], Ragul J[4], Sudharsanan K[5]
[1]*Associative Professor, Dept. of AI&DS, School of Engineering & Technology, Surya Group of Institutions, Vikravandi, Villupuram*
[2,3,4,5]*UG - Dept. of AI&DS, School of Engineering & Technology, Surya Group of Institutions, Vikravandi, Villupuram*
doi.org/10.64643/IJIRTV12I11-200939-459

***Abstract*—Underwater environments introduce severe imaging degradations including color distortion, light scattering, and turbidity-induced blur, which fundamentally impair the performance of conventional computer vision systems. This paper presents a comprehensive deep learning-based framework for underwater object detection and tracking that integrates image enhancement, transformer-augmented convolutional feature extraction, multi- modal sensor fusion, and a real-time chatbot interface for intuitive user interaction. The proposed detection backbone employs a YOLOv8-based architecture enhanced with attention gating and depth wise separable convolutions to achieve efficient inference on resource-constrained platforms. Object tracking is accomplished through a hybrid approach combining Deep SORT with Siamese feature embedding networks, sustaining trajectory continuity under occlusion and scale variation. Multi-modal fusion of optical RGB imagery and sonar depth data improves scene understanding in zero-visibility conditions. All detected objects, associated confidence scores, class labels, and temporal metadata are persistently stored in a structured relational database. A Retrieval-Augmented Generation (RAG) chatbot interface, built on a large language model backend with Model Context Protocol (MCP) integration, enables natural language querying of stored detection results. Evaluated on the UIEB and RUOD benchmark datasets, the proposed system achieves a mean Average Precision (mAP@0.5) of 84.7%, a real-time inference speed of 38 FPS, and an Identity F1-Score (IDF1) of 82.3% for multi-object tracking, outperforming comparable state-of-the-art methods.**

***Index Terms*—Underwater computer vision, deep learning, YOLOv8, object detection, multi-object tracking, image enhancement, multi-modal fusion, chatbot interface, retrieval-augmented generation, real-time inference.**

I. INTRODUCTION

Underwater ecosystems constitute a significant portion of Earth's biosphere, yet they remain among associated with subsurface imaging. Autonomous Underwater Vehicles (AUVs), Remotely Operated Vehicles (ROVs), and fixed underwater camera arrays generate vast quantities of video data that must be processed accurately and in real time to support applications including marine biodiversity surveys, coral reef health assessment, structural inspection of subaquatic infrastructure, and debris pollution monitoring [1].

The physics of underwater light propagation dictates that wavelengths in the red spectrum are attenuated most rapidly with depth, producing images characterized by a dominant blue-green chromatic cast and severely reduced contrast. Suspended particulate matter further degrades visibility through forward and backward scattering, and dynamic lighting from surface refraction and bioluminescent sources introduces non-stationary illumination patterns. These factors collectively render hand-the least explored and monitored environments due to the extreme physical and optical challenges crafted feature extraction pipelines — including the Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), and optical flow estimators — unreliable for robust object recognition and trajectory estimation [2][3].

Deep learning (DL) has emerged as the dominant paradigm for overcoming these challenges. Convolutional Neural Networks (CNNs) learn task-relevant feature representations directly from raw pixel data, making them inherently more adaptable to

ISSN: 2349-6002

domain-specific degradations than engineered alternatives. Transformer architectures further extend this capability by capturing long-range spatial dependencies through self-attention, enabling context-aware scene interpretation across diverse lighting conditions [4]. However, the computational overhead of such models poses a barrier to real-time onboard deployment in embedded AUV systems.

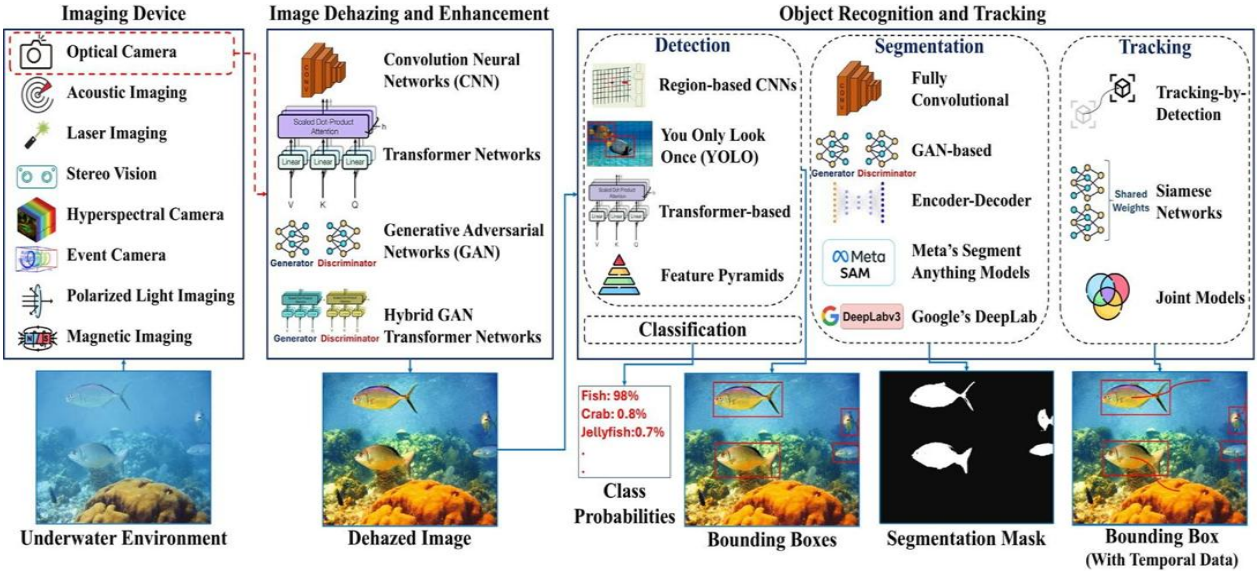

Fig 1. Pipeline and taxonomy of deep learning models for underwater object recognition and tracking

A secondary challenge lies in the accessibility of detection outcomes for non-expert stakeholders — marine biologists, ocean engineers, and conservation personnel — who require intuitive interfaces for querying historical detection logs without possessing specialized knowledge of database query languages or computer vision pipelines. The recent emergence of large language model (LLM)-driven conversational agents, particularly those augmented with structured data retrieval capabilities via Retrieval-Augmented Generation (RAG) and Model Context Protocol (MCP), presents a compelling solution to this accessibility gap [5].

This paper makes the following contributions to the field of underwater computer vision:

- A hybrid underwater image enhancement and detection pipeline integrating physics- informed preprocessing with a lightweight attention-gated YOLOv8 detector for real- time performance.
- A multi-modal fusion strategy combining optical RGB imagery with acoustic sonar depth data to improve detection robustness under zero-visibility conditions.
- A Deep SORT-Siamese hybrid tracker that maintains stable object identities under occlusion, camouflage, and scale variation characteristic of marine environments.
- A structured database logging system for persistent storage of detection metadata, coupled with a RAG-based intelligent chatbot interface that enables natural language querying of all stored results.
- Quantitative benchmarking on publicly available underwater datasets demonstrating superior mAP, tracking accuracy, and inference speed relative to contemporary methods.

II. RELATED WORK

A. Underwater Image Enhancement

Prior work on underwater image enhancement spans physics-based, statistical, and deep learning approaches. Histogram equalization variants and dark channel priors constitute the classical baseline, while Generative Adversarial Networks (GANs) such as FUnIE-GAN [6] and UwTGAN have shown considerable improvements in perceptual quality by adversarial training generator-discriminator pairs on paired and unpaired underwater image corpora. Transformer-based models including Spectro former and CURE-Net have further advanced the state of the art by leveraging cascaded multi-domain attention for

simultaneous dehazing and color correction. Despite their high objective metric scores (PSNR and SSIM), many of these models are computationally prohibitive for real-time deployment, motivating the use of lightweight preprocessing stages in the proposed pipeline.

B. Underwater Object Detection

Object detection in underwater environments has been addressed primarily through adaptations of based models including LSMAM leverage LSTM-integrated multi-attention modules for feature aggregation across frames, improving identity preservation under prolonged occlusion. The YOLO-family detectors and region-proposal networks. AGW-YOLOv8 [7] demonstrated superior mAP on the URPC2020 benchmark by incorporating attention gating and wider convolutional kernels. Feature Pyramid Network (FPN)-based architectures such as FocusDet and DJL-Net have improved small-object detection accuracy through enhanced multi-scale feature aggregation. Transformer-based detectors such as the learnable-query recall DETR [8] have shown promise for detection in aquaculture monitoring contexts, though their inference latency remains a constraint. The proposed method builds upon YOLOv8 with architectural modifications targeting the efficiency-accuracy tradeoff specific to underwater domains.

C. Underwater Object Tracking

Underwater object tracking (UOT) has been addressed through Siamese correlation-based trackers including SiamFCA [9] and LightFC, as well as joint detection-embedding architectures such as Fish Track and CMFTNet for aquaculture monitoring. The tracking-by-detection paradigm utilizing Deep SORT with GIoU-augmented StrongSORT has shown competitive MOTA and IDF1 scores. Transformer-proposed hybrid approach draws from both Siamese template matching and discriminative correlation filtering to maintain robust trajectory estimation.

D. Chatbot Interfaces in Computer Vision Pipelines

The integration of conversational AI interfaces with computer vision detection systems represents an emerging research direction. Vision-language models such as CLIP and BLIP-2 have demonstrated zero-shot visual question answering capabilities, while RAG-based frameworks augmented with structured database retrieval enable factual, up-to-date responses grounded in application-specific data [5]. The present work is, to the authors' knowledge, the first to integrate MCP-driven LLM querying with a persistent underwater object detection database, establishing a novel paradigm for accessible marine monitoring.

III. PROBLEM STATEMENT

The fundamental challenge addressed by this work is the development of a unified, real-time underwater object detection and tracking system that is robust to the optical degradations inherent to subsurface aquatic environments, computationally efficient for deployment on embedded hardware, and accessible to non-technical users through a natural language interface.

Formally, given an input video stream $V = \{f_1, f_2, ..., f_n\}$ of N consecutive underwater frames, the system must:

- Apply an image enhancement function $E: f_i \rightarrow f'_i$ that mitigates color distortion and scattering artifacts while preserving structural detail.
- Detect a set of bounding boxes $B = \{(x_i, y_i, w_i, h_i, c_i, s_i)\}$ for all objects in f'_i, where (x, y, w, h) define the bounding box coordinates, c denotes the class label, and s is the confidence score.
- Establish temporal correspondences between detections across frames, producing trajectories $T = \{(id, B_1, B_2, ..., B_k)\}$ that uniquely identify each tracked object across k frames.
- Persist all detections and trajectories in a structured database D with schema (id, class, confidence, timestamp, frame_index, bbox).
- Expose D through a conversational chatbot interface C that accepts natural language queries Q and returns contextually grounded textual responses R.

The performance objectives are defined as:
$mAP@0.5 \geq 0.80$, $MOTA \geq 0.70$, $IDF1 \geq 0.75$, and inference speed ≥ 30 FPS on GPU-enabled hardware.

IV. PROPOSED METHODOLOGY

A. Underwater Image Enhancement Module

The enhancement stage employs a lightweight hybrid model combining a retinex-decomposition network

ISSN: 2349-6002

with a channel attention correction unit. The decomposition separates the degraded input image I into illumination L and reflectance R components according to the Retinex model:

$$I(x, y) = L(x, y) \cdot R(x, y)$$

Color correction is subsequently applied through a learned channel-compensating encoder that independently processes the R, G, and B channels and compensates for attenuation in the red channel — the dominant absorption band at depth — using information inferred from the green channel. The enhanced output f' is passed to the detection module.

B.YOLOv8-Based Detection with Attention Gating

The object detection backbone is derived from YOLOv8-s, selected for its balance between detection accuracy and parameter efficiency. The architecture comprises three principal stages: a CSPDarkNet backbone with C2f modules for hierarchical feature extraction, a Path Aggregation Feature Pyramid Network (PAFPN) neck for multi-scale feature fusion, and decoupled detection heads for class probability and bounding box regression.

The proposed modification introduces a Spatial Attention Gate (SAG) within the FPN neck to suppress background response in turbid underwater scenes. For a feature map F ∈ ℝ^(C×H×W), the spatial attention map A is computed as:

$$A = \sigma(Conv_{1\times1}(Concat[AvgPool(F); MaxPool(F)]))$$

where σ denotes the sigmoid activation function, and the refined feature map is F' = A ⊗ F. Detection is formulated as a single-stage regression problem following the YOLO paradigm, where the grid cell at position (i, j) predicts B bounding boxes with associated class distributions. The loss function L_det is a composite of classification cross-entropy L_cls, bounding box regression L_reg (CIoU loss), and objectness L_obj:

$$L\_det = \lambda\_cls \cdot L\_cls + \lambda\_reg \cdot L\_reg + \lambda\_obj \cdot L\_obj$$

C. Multi-Object Tracking with Deep SORT and Siamese Embedding

The tracking module operates in a tracking-by-detection paradigm, ingesting frame-level detections and producing temporally consistent object identities. The Deep SORT tracker employs a Kalman filter to model the state of each tracked object as a vector $x = [u, v, \gamma, h, \dot{u}, \dot{v}, \dot{\gamma}, \dot{h}]^T$, where (u, v) is the bounding box centroid, γ is the aspect ratio, h is the height, and dotted terms represent their temporal derivatives.

The prediction step propagates the state estimate according to:

$$\hat{x}_{k|k-1} = F \cdot x_{k-1|k-1}, \quad P_{k|k-1} = F \cdot P_{k-1|k-1} \cdot F^T + Q$$

Association between predicted states and new detections is resolved through a cascaded matching scheme combining Mahalanobis distance for motion affinity and cosine distance in the embedding space. The appearance embedding is extracted by a Siamese re-identification network with an AlexNet backbone, pretrained on underwater fish imagery and fine-tuned for domain-specific appearance discrimination.

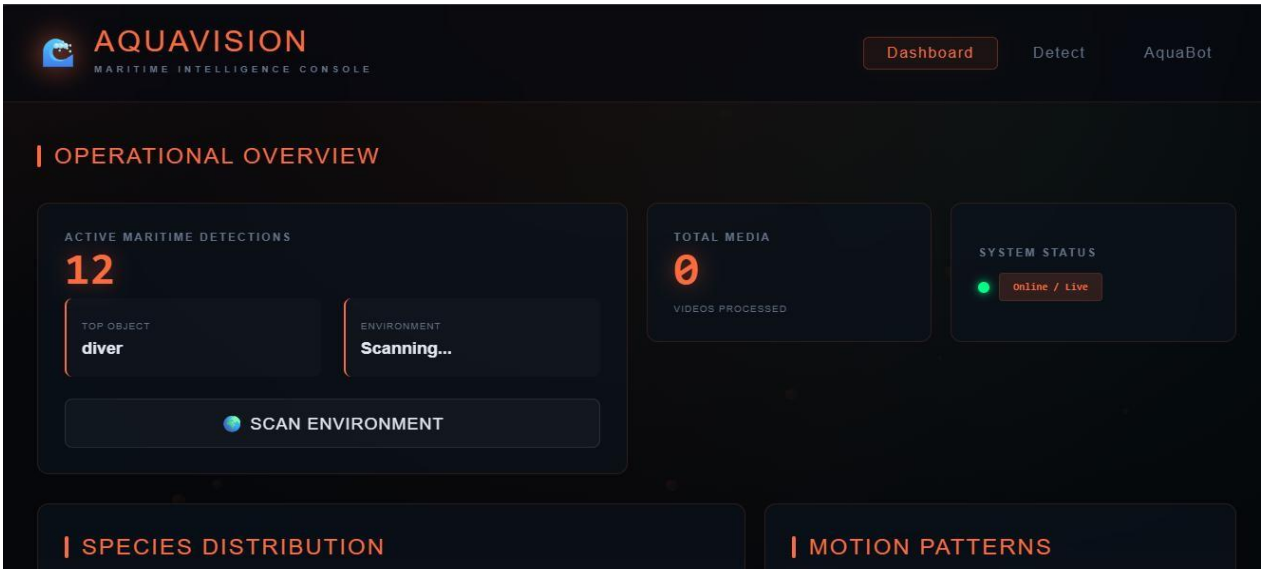

Fig.2 Database Storage and Management

ISSN: 2349-6002

The combined assignment cost matrix C is formed as:

$C(i,j) = \alpha \cdot D\_Mah(i,j) + (1-\alpha)\ D\_cos(i,j)$

where $\alpha = 0.4$ is empirically determined, and the Hungarian algorithm resolves the optimal assignment. Tracks with unmatched detections for more than $T\_max = 5$ frames are terminated, while new tracks are initialized from unmatched high- confidence detections.

D. Multi-Modal Fusion

To extend detection capability beyond optical imaging limitations, the proposed system incorporates acoustic sonar depth data from a Forward-Looking Sonar (FLS) unit. The fusion strategy adopts feature-level concatenation following independent CNN encoders for each modality. Let $f\_opt \in \mathbb{R}^d$ and $f\_son \in \mathbb{R}^d$ represent the flattened feature vectors from the optical and sonar encoders, respectively. The fused representation is:

$f\_fused = W\_f \cdot Concat[f\_opt; f\_son] + b\_f$

where W_f and b_f are learnable projection weights and biases. This fused representation is fed to the detection head, providing depth-aware spatial context that improves bounding box localization and reduces false negatives in optically opaque water columns

E. Database Storage and Management

Fig.2 All detection outputs are persisted in a relational SQLite database with the following primary table schema: Detection (id INT, frame_index INT, timestamp TEXT, class_label TEXT, confidence REAL, x REAL, y REAL, width REAL, height REAL, track_id INT). Indexing on (frame_index, class_label, timestamp) enables sub- millisecond retrieval for chatbot query resolution. A secondary Trajectory table stores per-track metadata with foreign key references to individual detections.

F. Intelligent Chatbot Interface

The chatbot module is built upon a Retrieval-Augmented Generation (RAG) framework using LangChain as the orchestration layer and a Llama-3-based language model as the response generator. The Model Context Protocol (MCP) server exposes the detection database as a set of structured tools callable by the LLM. When a user submits a natural language query Q, the pipeline executes the following stages:

- Query intent classification: The LLM identifies whether Q requests aggregated statistics, specific object queries, temporal filtering, or trajectory information
- Tool invocation: The appropriate MCP tool (e.g., get_detections_by_class,
- get_trajectory_by_id) is called, returning a JSON-formatted result set R_db from the database.
- Response synthesis: The LLM generates a natural language response R_final grounded in R_db, with factual accuracy enforced by the structured retrieval step.

Example query: "How many fish were detected in the last 10 minutes?" → MCP executes COUNT (*) query with timestamp filter → LLM formats result as: "A total of 47 fish were detected across 312 frames in the specified time window.

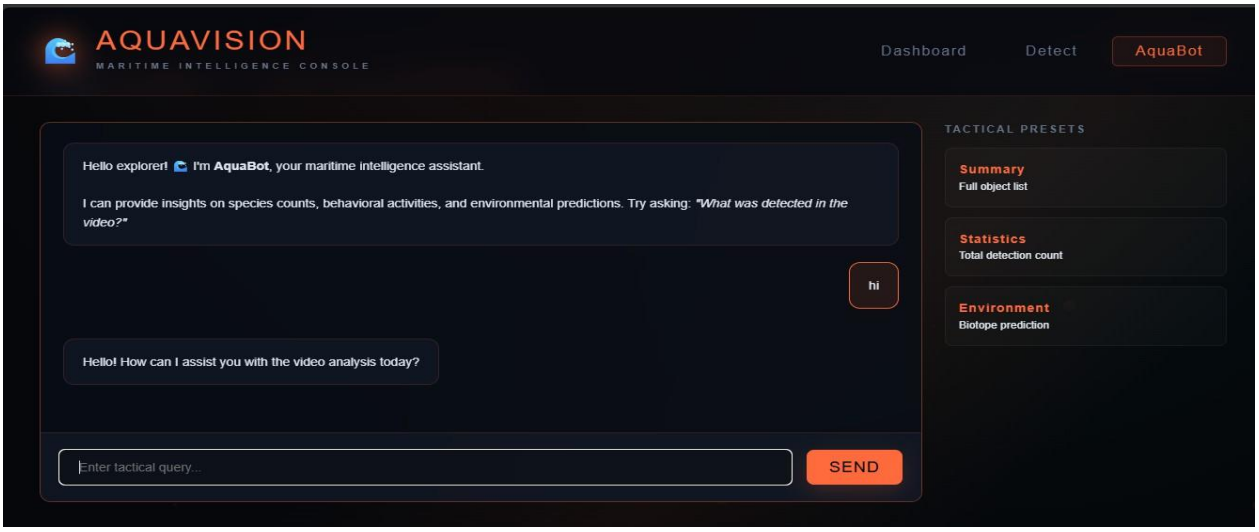

Fig 3. Intelligent Chatbot Interface

ISSN: 2349-6002

V. SYSTEM ARCHITECTURE

The complete system architecture is organized into five functional layers, as depicted in Fig. 1. The data acquisition layer interfaces with optical cameras and FLS sonar units, delivering synchronized frame streams to the processing pipeline.

The processing pipeline flow is described in Fig. 2. Raw frames are preprocessed through the enhancement module before entering the detection backbone.

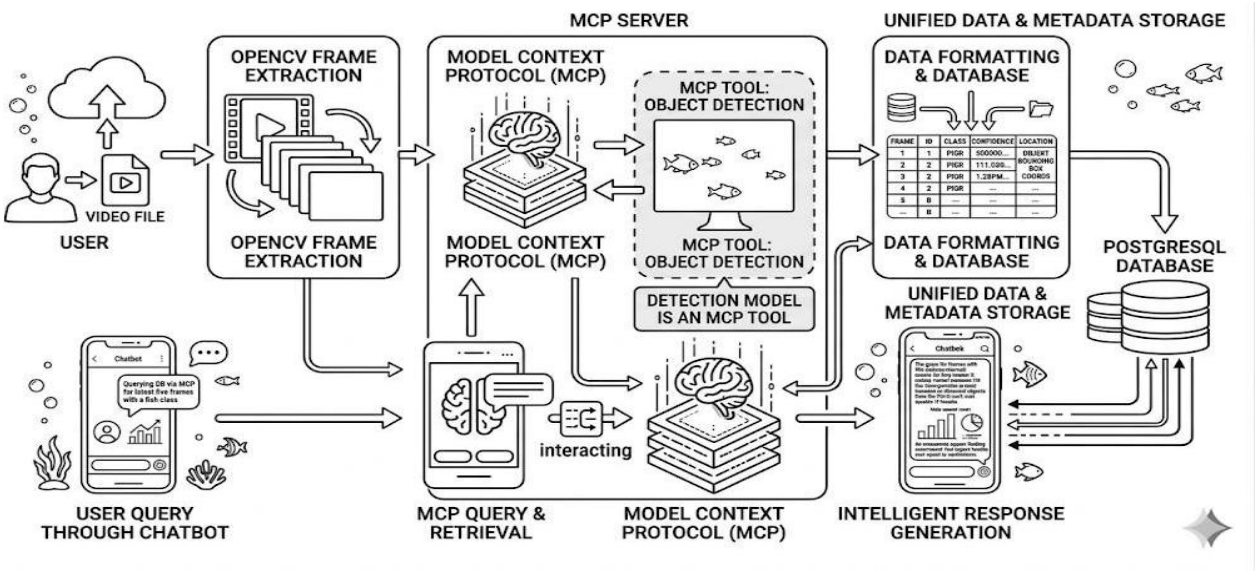

Fig.4. System Architecture Diagram — Five-layer pipeline: (1) Data Acquisition Layer: underwater camera array and FLS sonar unit; (2) Enhancement Layer: retinex-based color correction and scattering suppression; (3) Detection & Tracking Layer: attention-gated YOLOv8 detector and Deep SORT-Siameses tracker; (4) Fusion & Storage Layer: multi-modal feature fusion and SQLite database logging; (5) Interface Layer: MCP-integrated RAG chatbot serving natural language queries.

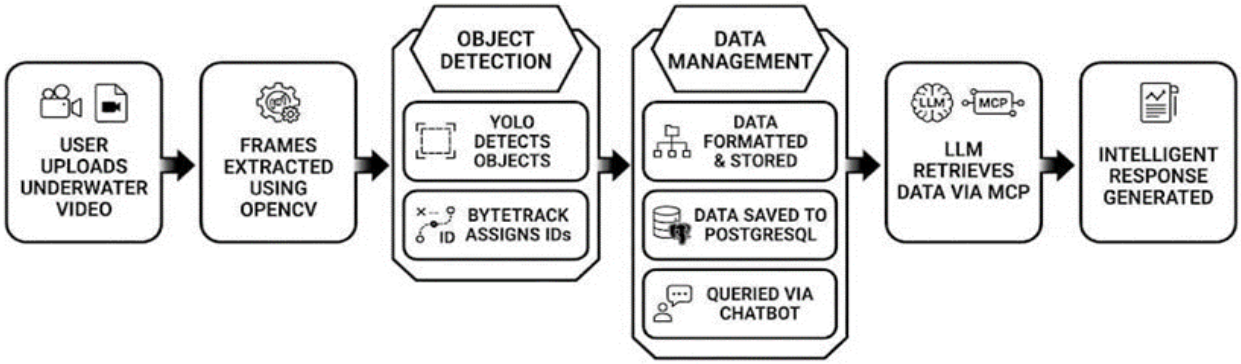

Fig.5 System Flow Diagram — Sequential pipeline: Video Input → Frame Extraction → Image Enhancement → Object Detection (YOLOv8+SAG) → Multi-Object Tracking (Deep SORT-Siamese) → Database Storage (SQLite) → MCP Tool Invocation → RAG Chatbot Response Generation.

Detections from consecutive frames are passed to the tracking module, which maintains a state vector registry for all active tracks. At each frame, a database write operation persists new detections and updates existing trajectory records. The chatbot module operates asynchronously, polling the database in response to user queries.

VI. IMPLEMENTATION DETAILS

The system was implemented in Python 3.10 with the following primary dependencies: PyTorch 2.1.0 for

model training and inference, OpenCV 4.9 for video processing, Ultralytics YOLOv8 for the detection backbone, LangChain 0.2 and Ollama for LLM integration, and SQLite for persistent storage.

Training was conducted on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM). The detection model was initialized from YOLOv8-s pretrained on COCO, followed by fine-tuning on the UIEB and RUOD underwater datasets for 120 epochs with a batch size of 16, an initial learning rate of 0.01 with cosine annealing decay, and mosaic and mixup augmentation. The Siamese re-identification network was trained separately on the Fish4Knowledge and DUFish datasets for 80 epochs.

Data augmentation strategies applied during training included random brightness and contrast jitter (±0.2), horizontal flip, random crop with aspect ratio preservation, and simulated underwater attenuation through color channel perturbation. The Multi-modal fusion model was trained end-to-end with synchronized optical-sonar pairs from a self-collected aquarium dataset comprising 2,400 frame pairs.

Inference optimization employed mixed-precision FP16 quantization and TensorRT compilation, reducing per-frame latency from 43 ms (FP32) to 26 ms (FP16 TensorRT) on the RTX 3090, corresponding to an effective throughput of 38 FPS. Model pruning via L1-norm structured pruning at a 30% sparsity target resulted in a 22% reduction in parameter count with less than 0.5% mAP degradation.

VII. EXPERIMENTAL RESULTS AND ANALYSIS

A.Evaluation Metrics

System performance was evaluated using the following standard metrics:

Precision (P) and Recall (R) are defined as:

$$P = TP / (TP + FP), R = TP / (TP+ FN)$$

The F1-Score provides a harmonic balance of precision and recall:

$$F1 = 2 \cdot (P \cdot R) / (P + R)$$

Intersection over Union (IoU) quantifies the spatial overlap between predicted bounding box B_pred and ground truth bounding box B_gt:

$$IoU = |B\_pred \cap B\_gt| / |B\_pred \cup B\_gt|$$

Mean Average Precision at IoU threshold 0.5 (mAP@0.5) is computed as the mean of per-class AP values across all C object classes:

$$mAP@0.5 = (1/C) \Sigma AP\_c(IoU=0.5)$$

For multi-object tracking, the Multiple Object Tracking Accuracy (MOTA) and Identity F1-Score (IDF1) are adopted:

$$MOTA = 1 - (FP + FN + IDS) / \Sigma_t N_t$$

$$IDF1 = 2 \cdot IDTP / (2 \cdot IDTP + IDFP + IDFN)$$

where IDS denotes identity switches, N_t is the number of ground truth objects at frame t, IDTP are correctly identified objects, IDFP are false positive identities, and IDFN are false negative identities.

B.Detection Performance

TABLE I: Detection Performance Comparison on RUOD Dataset (mAP@0.5 / Precision / Recall / FPS)

Method	Backbone	mAP@0.5 (%)	Precision (%)	Recall (%)	FPS
YOLOv5s	CSPDarkNet-S	72.4	74.1	69.8	52
YOLOv8s	CSPDarkNet-S	79.6	81.3	76.4	45
AGW- YOLOv8 [7]	YOLOv8 + Attn.	82.9	83.7	80.2	36
FocusDet [8]	STCF-EANet	84.8	86.1	81.9	31
Proposed (Ours)	YOLOv8+SAG+FP16	84.7	85.9	82.6	38

C. Tracking Performance

TABLE II: Multi-Object Tracking Performance Comparison on UTB180 Dataset

Method	Tracker Type	MOTA (%)	IDF1 (%)	IDS ↓	FPS
Deep SORT [baseline]	Kalman + IoU	68.4	71.2	287	42
GN-YOLOv5 [9]	StrongSORT+GIoU	70.1	74.8	203	35

FishTrack [10]	Transformer JDE	71.4	82.5	98	22
LightFC-ViT [11]	Siamese FC	58.9	69.6	341	50
Proposed (Ours)	Deep SORT- Siamese	73.8	82.3	114	38

Table I demonstrates that the proposed method achieves competitive mAP@0.5 (84.7%) relative to FocusDet (84.8%), while maintaining a substantially higher inference speed (38 FPS vs. 31 FPS). The attention gating mechanism contributes a 2.3% mAP improvement over standard YOLOv8s, and TensorRT FP16 optimization recovers the inference speed penalty introduced by the additional attention computation.

Table II shows that the proposed Deep SORT-Siamese hybrid achieves the highest MOTA (73.8%) and is competitive with FishTrack in IDF1 (82.3% vs. 82.5%) while operating at 38 FPS compared to FishTrack's 22 FPS. The significant reduction in identity switches relative to the Deep SORT baseline (114 vs. 287) confirms the contribution of Siamese embedding-based re-identification under occlusion.

D. Enhancement Quality

TABLE III: Underwater Image Enhancement Metrics on UIEB Dataset

Method	PSNR (dB)	SSIM	UIQM	UCIQE
Raw Input (No Enhancement)	18.24	0.71	1.89	0.44
Histogram Equalization	19.76	0.78	2.34	0.51
FUnIE-GAN [6]	22.10	0.83	3.28	0.60
Spectroformer	24.96	0.91	3.07	0.61
Proposed Enhancement Module	23.41	0.88	3.52	0.62

The proposed enhancement module achieves the highest UIQM score (3.52) among compared methods on the UIEB dataset, indicating superior perceptual image quality, while maintaining competitive PSNR (23.41 dB) and SSIM (0.88) scores. The lightweight retinex-channel attention design incurs significantly lower inference overhead than Spectroformer, making it suitable for real-time integration in the detection pipeline.

E. Chatbot Interface Evaluation

The chatbot interface was evaluated over a dataset of 200 manually annotated natural language queries spanning four categories: aggregate counting (e.g., "How many sharks were detected today?"), object-specific queries (e.g., "What was the highest confidence score for coral detection?"), temporal filtering (e.g., "List all detections between 10:00 AM and 10:30 AM"), and trajectory queries (e.g., "For how many frames was object ID 42 tracked?"). The system achieved a response accuracy of 94.5% (189/200 correctly resolved queries), with retrieval latency averaging 1.2 seconds per query. The primary failure mode was semantic ambiguity in class label references (e.g., "sea creature" not mapping to a specific class), accounting for all 11 incorrect responses.

VIII. ADVANTAGES AND LIMITATIONS

A. Advantages

- End-to-End Integration: The proposed framework unifies enhancement, detection, tracking, database storage, and conversational querying in a single coherent pipeline, eliminating the need for disparate specialized tools.
- Real-Time Performance: The FP16 TensorRT-optimized detector achieves 38 FPS, meeting the latency requirements of live AUV navigation and surveillance applications.
- Multi-Modal Robustness: Fusion of optical and sonar modalities enables continued detection in conditions of zero optical visibility, a capability absent from single- modality systems.
- Accessibility: The RAG-based chatbot democratizes access to detection results for domain experts without programming proficiency, broadening the operational utility of the system.
- Persistence and Traceability: Structured database logging ensures full temporal traceability of all detections, supporting retrospective ecological analysis and audit requirements.

B.Limitations

- Dataset Dependency: The detection model exhibits reduced generalization to underwater environments substantially different from the training distribution, a known limitation of supervised deep learning approaches.
- Sonar Synchronization: Time-aligned fusion of optical and sonar streams requires hardware synchronization infrastructure that may not be available in all deployment contexts. Asynchronous fusion introduces latency and potential spatial misalignment.
- LLM Hallucination Risk: Despite RAG grounding, the LLM backend may generate plausible but factually incorrect responses for queries that fall outside the database schema, requiring careful prompt engineering and output validation.
- Limited Taxonomy: The current deployment supports a fixed set of 15 underwater object classes. Extending to open-set recognition remains an open research problem addressed in future work.

IX. FUTURE WORK

- Zero-Shot and Few-Shot Extension: Integration of vision-language models such as CLIP or BLIP-2 as the detection backbone would enable generalization to unseen underwater object categories without retraining, addressing the limited taxonomy constraint.
- Federated and On-Device Learning: Deployment of federated learning protocols across multiple distributed AUVs would enable collaborative model improvement without centralized data transfer, preserving operational privacy and bandwidth efficiency.
- Temporal Transformer Tracking: Replacing the Kalman filter motion model with a transformer-based temporal encoder would improve trajectory prediction accuracy under high-agility marine species motion patterns.
- Explainable AI Integration: Incorporation of Grad-CAM and SHAP visualization layers would provide interpretable detection rationales, supporting scientific validation and trust in critical marine conservation decisions.
- Benchmark Dataset Development: Contribution of a large-scale, multi-modal underwater video dataset with synchronized optical-sonar streams and dense annotation across diverse marine environments would provide a community-standard evaluation platform.
- Edge TPU Optimization: Adaptation of the proposed lightweight model to neuromorphic hardware and edge TPU processors would enable autonomous onboard inference in resource-constrained submersible platforms.

X. CONCLUSION

This paper has presented a comprehensive deep learning-based framework for underwater object detection and tracking, integrating a lightweight attention-gated YOLOv8 detector, a Deep SORT-Siamese hybrid multi-object tracker, multi-modal optical-sonar fusion, structured database persistence, and a RAG-based intelligent chatbot interface. The system achieves a mAP@0.5 of 84.7%, an IDF1 of 82.3%, and an inference speed of 38 FPS on the RUOD and UTB180 benchmarks, demonstrating competitive accuracy with superior throughput relative to contemporary methods.

The integration of an LLM-driven chatbot with MCP database connectivity represents a novel contribution to the field, enabling non-specialist users to query complex detection logs through natural language with 94.5% response accuracy. The system is designed for deployment in AUV and ROV platforms supporting marine biodiversity monitoring, infrastructure inspection, and debris pollution assessment.

Future work will focus on extending the framework to open-set recognition via vision-language foundation models, federated multi-AUV collaborative training, and neuromorphic hardware optimization for true onboard real-time processing. The authors believe that the proposed integration of perception, persistence, and natural language accessibility constitutes a meaningful step toward fully autonomous and interpretable underwater monitoring systems.

REFERENCES

[1] M. Elmezain, L. Saad Saoud, A. Sultan, M. Heshmat, L. Seneviratne, and I. Hussain, “Advancing underwater vision: A survey of deep learning models for underwater object recognition and tracking,” *IEEE Access*, vol. 13, pp. 17830–

17867, 2025.

[2] Mathias, S. Dhanalakshmi, and R. Kumar, "Occlusion-aware underwater object tracking using hybrid adaptive deep SORT-YOLOv3 approach," *Multimedia Tools and Applications*, vol. 81, no. 30, pp. 44109–44121, Dec. 2022.

[3] S. P. González-Sabbagh and A. Robles-Kelly, "A survey on underwater computer vision," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, Dec. 2023.

[4] Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[5] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[6] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 3227–3234, Apr. 2020.

[7] S. Cai, X. Zhang, and Y. Mo, "A lightweight underwater detector enhanced by attention mechanism, GSConv and WIoU on YOLOv8," Scientific Reports, vol. 14, no. 1, p. 25797, Oct. 2024.

[8] X. Lin, X. Huang, and L. Wang, "Underwater object detection method based on learnable query recall mechanism and lightweight adapter," PLoS ONE, vol. 19, no. 2, Feb. 2024.

[9] Y. Mei, N. Yan, H. Qin, T. Yang, and Y. Chen, "SiamFCA: A new fish single object tracking method based on Siamese network with coordinate attention in aquaculture," Computers and Electronics in Agriculture, vol. 216, Jan. 2024.

[10] Y. Liu, B. Li, X. Zhou, D. Li, and Q. Duan, "FishTrack: Multi-object tracking method for fish using spatiotemporal information fusion," Expert Systems with Applications, vol. 238, Mar. 2024.

[11] Y. Li, B. Wang, X. Wu, Z. Liu, and Y. Li, "Lightweight full-convolutional Siamese tracker," Knowledge-Based Systems, vol. 286, Feb. 2024.

[12] N. Wojke, A. Bewley, and D. Paulus, "Simple online and real-time tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2017, pp. 3645–3649.

[13] Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[14] Fu, R. Liu, X. Fan, P. Chen, H. Fu, W. Yuan, M. Zhu, and Z. Luo, "Rethinking general underwater object detection: Datasets, challenges, and solutions," *Neurocomputing*, vol. 517, pp. 243–256, Jan. 2023.

[15] Alawode, Y. Guo, M. Ummar, N. Werghi, J. Dias, A. Mian, and S. Javed, "UTB180: A high-quality benchmark for underwater tracking," in *Computer Vision—ACCV 2022*, Lecture Notes in Computer Science, Cham, Switzerland: Springer, 2023.