Survey | Received 18 November 2025; Revised 24 February 2026; Accepted 19 March 2026; Published 16 June 2026
https://doi.org/10.55092/rl20260020

Deep learning for underwater object detection: a comprehensive survey of models, datasets, and challenges

Hari Bhandari* and Pengcheng Liu*

Department of Computer Science, University of York, York, UK

* Correspondence authors; E-mails: hb1622@york.ac.uk (H.B.); pengcheng.liu@york.ac.uk (P.L.).

Highlights:

- Surveys deep learning approaches for underwater object detection, including CNN, Transformer, and hybrid models.
- Reviews major underwater imaging challenges and their impact on detection performance.
- Summarises key underwater datasets, benchmarks, and evaluation metrics.
- Examines enhancement–detection pipelines and common failure modes.
- Identifies open research challenges and future directions for underwater vision.

Abstract: This survey provides a comprehensive synthesis of methods, datasets, metrics, and deployment strategies from the evolution of convolutional neural network (CNN)-based detectors to emerging transformer and hybrid architectures. It unifies fragmented literature into a structured taxonomy while integrating results from 2014–2025 studies. The paper reviews benchmark datasets, discusses evaluation protocols and reproducibility standards, and proposes a deployment playbook considering latency, energy, and hardware constraints. Beyond technical performance, it addresses responsible AI practices and ethical challenges in marine observation. By highlighting open problems in multimodal fusion, self-supervised learning, and on-device adaptation, this work aims to guide future research and practical deployment of underwater vision systems. A comprehensive survey of underwater object detection covering classic CNN-based detectors, modern transformer and hybrid models, training and evaluation practices under challenging aquatic conditions, the dataset landscape, deployment constraints (latency/VRAM/energy), and open problems for real-world marine applications.

Keywords: underwater object detection; YOLO; RT-DETR; transformers; turbidity; dataset bias; edge deployment; mAP; VRAM; marine robotics

1. Introduction

1.1. Context and significance

The study of underwater object detection (UOD) is critical for biodiversity conservation, sustainable fisheries management, and the early detection of environmental hazards. A widely cited observation

Copyright©2026 by the authors. Published by ELSP. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

is that we have higher-resolution maps of Mars than of much of Earth's seafloor [1]. Oceans cover approximately 71% of Earth's surface [2]. Effective observation is essential for protecting endangered species, setting evidence-based fishing limits, and warning coastal communities of risks. Traditional survey methods, such as diver transects, static cameras, and heuristic image filters, struggle when visibility drops, lighting changes, or organisms hide within seafloor clutter. While artificial intelligence offers continuous, non-intrusive surveillance, the community lacks decisive comparisons of the newest detector families. Existing surveys rarely provide head-to-head comparisons across convolutional neural network (CNN), transformer, and hybrid detectors while also considering underwater imaging physics and embedded deployment constraints. Without such comparisons, practitioners can end up deploying inefficient or inaccurate models, burning through limited power and bandwidth under conditions that are already demanding.

1.2. Motivation

Marine biodiversity is increasingly threatened by climate change, pollution, and overfishing. These pressures hit ecosystems and coastal communities hard, particularly those whose livelihoods depend on fisheries and tourism [3,4]. As marine habitats deteriorate and species numbers drop, these communities face economic hardship. Effective monitoring informs decisions about marine protected areas, sustainable fishing quotas, and aquaculture management. Shrimp/prawn farms, for example, depend on timely stock monitoring (e.g., size/weight estimation and feeding-related observations) to reduce mortality and optimise feeding regimes [5–8].

Monitoring marine life is still far from straightforward. Offshore platforms typically rely on battery-powered devices with limited computing capabilities, similar to portable instruments used during remote fieldwork. Modern detectors, particularly CNNs, have improved underwater object detection. Even so, there is no clear side-by-side comparison of the latest CNN, transformer-based, and hybrid detection architectures. Without solid evidence, practitioners risk choosing inefficient or inaccurate models, limiting their effectiveness in real-world marine monitoring.

1.3. Aim of the study

This survey contributes: (1) an analysis of recent studies on CNN-, Transformer-, and hybrid-based detection architectures; (2) broadened coverage of datasets and benchmark challenges, including a comparative analysis of new large-scale underwater datasets; and (3) a discussion of evaluation metrics, emerging trends, and open research challenges. This survey consolidates the evidence base on underwater object detection by comparing how convolutional, transformer-based, and hybrid architectures cope with domain-specific constraints. It emphasises four objectives: (1) distilling the imaging physics that dominate performance limits, (2) cataloguing datasets and benchmark challenges that span coastal, pelagic, and aquaculture settings, (3) analysing architectural trends—from classic two-stage CNNs through lightweight You Only Look Once (YOLO) variants to multimodal hybrids—with respect to accuracy, efficiency, and robustness, and (4) highlighting evaluation protocols, metrics, and open challenges that prevent routine deployment on autonomous platforms.

1.4. Scope of the work and approach

The review covers peer-reviewed and preprint literature published between 2014 and 2025 across IEEE Xplore, Elsevier, arXiv, and major robotics proceedings. Inclusion criteria required explicit underwater benchmarks or ablation studies that account for turbidity, colour attenuation, or embedded-compute limits. Each paper was coded for dataset usage, architectural family, enhancement pipeline, and reported metrics, enabling cross-study comparison without reproducing project-specific implementations. This methodology allows the manuscript to focus on synthesising insights from the broader field rather than detailing specific experiments.

Although centred on still-image detection, the analysis provides a foundation for adjacent tasks such as underwater tracking, multimodal data fusion, and bioacoustic integration. The paper is structured as follows: Section 2 discusses the methodology adopted to conduct the literature survey. Section 3 examines underwater imaging challenges, Section 4 surveys datasets and benchmarks, Section 5 categorises deep learning approaches, Section 6 discusses enhancement-plus-detection pipelines, Section 7 reviews evaluation metrics, Section 8 outlines open challenges and emerging research directions, and Section 9 concludes with key recommendations for the community.

2. Survey methodology

This survey synthesises literature on deep learning for underwater object detection published between 2014 and 2025. To ensure comprehensive and reproducible coverage, we adopted a systematic search and coding protocol.

2.1. Literature search protocol

We queried four major databases IEEE Xplore, ACM Digital Library, arXiv, and Google Scholar using the following keyword combinations: (underwater or marine or aquatic) and (object detection or target detection or instance detection) and (deep learning or CNN or YOLO or transformer or neural network). The search covered publications from January 2014 through February 2025, capturing the evolution from early CNN-based detectors to recent transformer and hybrid architectures.

2.2. Inclusion and exclusion criteria

Studies were included if they: (1) presented empirical results on underwater object detection tasks, (2) employed deep learning methods, (3) reported quantitative metrics (mean Average Precision (mAP), precision, recall, or tracking scores), and (4) were published in peer-reviewed venues or as validated preprints with reproducible experiments. We excluded purely theoretical papers without experimental validation, non-English publications, and works focused exclusively on image enhancement or segmentation without detection components.

2.3. Screening and coding

Initial keyword searches returned 487 candidate papers. After title and abstract screening, 214 papers met the inclusion criteria. Full-text review further refined this to 156 papers that form the core of this

survey. Each paper was coded for: (1) dataset usage (e.g., Underwater Robot Picking Contest (URPC), Brackish, FathomNet), (2) architectural family (CNN, YOLO, Detection Transformer (DETR), hybrid), (3) enhancement pipeline (if any), and (4) reported metrics (mAP, Frames Per Second (FPS), Video Random Access Memory (VRAM), tracking scores). This structured coding enables the quantitative comparisons and taxonomic analyses presented in subsequent sections.

3. Underwater imaging challenges

Underwater object detection systems face a complex photometric transfer function before a single pixel reaches the network. Light rapidly attenuates as wavelengths travel through water, causing long-wavelength reds to disappear within the first few metres and leaving a blue-green dominant spectrum [9]. Suspended particulates then scatter the remaining light, producing backscatter haze and non-uniform veiling glare that reduce contrast and blend foreground organisms into the background [10]. These effects vary with depth, salinity, and solar angle, so an image captured minutes apart at the same site can present markedly different colour statistics, complicating any effort to learn stable priors [11–13].

3.1. Environmental variability and scene complexity

Even mild turbidity introduces marine snow streaks, caustic highlights, and depth-dependent blurring that upset traditional feature detectors and modern convolutional filters alike [14]. Benthic habitats add clutter from corals, rocks, or man-made debris whose textures mimic target species, while aquaculture cages introduce nets and biofouling that occlude small fish. Biological activity amplifies this complexity: swarming sardines, translucent jellyfish, and overlapping crustaceans yield dense scenes with frequent partial occlusion and non-rigid deformation, all of which degrade localisation accuracy if detectors rely solely on high-frequency cues.

3.2. Sensor and platform constraints

Field deployments seldom enjoy unconstrained sensing. Battery-powered autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs) must manage limited power budgets, throttled bandwidth, and vibrations introduced by thrusters. Hardware therefore ranges from compact rolling-shutter Red Green Blue (RGB) modules to dual-modality optical-sonar rigs that need cross-sensor synchronisation [15,16]. In shallow coastal zones, surface reflections saturate sensors, whereas deep deployments rely on strobes that create specular hotspots and hard shadows. These factors demand models that can adapt across camera intrinsics, frame rates, and noise regimes without per-mission retraining.

3.3. Implications for algorithm design

The combination of optical degradation, environmental clutter, and hardware constraints leads to pronounced domain shift relative to terrestrial datasets such as Common Objects in Context (COCO) or PASCAL Visual Object Classes (VOC). Without explicit compensation, detectors overfit to specific water types or camera rigs, yielding brittle behaviour when moved to new reefs or farms [17,18]. Recent studies emphasise unsupervised or weakly supervised adaptation, domain-aware augmentation, and multimodal fusion as practical responses. However, reproducible pipelines remain scarce because many datasets provide limited

coverage of depth, turbidity, or illumination permutations [18]. Documenting these challenges explains why the subsequent sections devote equal emphasis to datasets, enhancement strategies, and evaluation metrics alongside architectural advances.

4. Datasets and benchmark challenges

Reliable conclusions about underwater detection hinge on datasets that capture spectral distortion, species diversity, and deployment constraints. Early studies often reused terrestrial corpora and added synthetic turbidity, but recent releases span brackish estuaries, coral reefs, pelagic habitats, and aquaculture cages. Table 1 summarises representative datasets and challenges that recur across contemporary literature.

Table 1. Representative underwater datasets and benchmark challenges with quality indicators.

Dataset	Modality	Images	Annotation(s)	Annotation(s) Quality	Class Balance	Depth (m)	Turbidity	Highlights
Brackish [19]	RGB stills	14,518	Boxes (6 multiple species)	Expert	Severe imbalance	∼9	High	Fixed-rig beneath Limfjords Bridge; strong turbidity gradients; class imbalance challenges
SUIM [20]	RGB segmentation	1525	Pixel masks (8)	Expert	Balanced	0–30	Mixed	Per-pixel cues for marine snow, reefs, fauna; used for enhancement pretraining
DeepFish [21]	RGB stills	∼40 k	Cls/Points/Seg	Crowd + expert	Moderate imbalance	2–40	Low–Med	20 Australian habitats; geotags; depth ranges; enables domain adaptation experiments
Fish4Knowledge [22]	Video	27 k clips	Boxes + tracks	Auto + manual	Long-tail	3–15	Low	Coral reef monitoring; long temporal sequences for tracking and behaviour analysis
FathomNet [23]	RGB stills	$>$ 84 k (growing)	Boxes + taxa	Crowd (variable)	Long-tail	0–4000	Mixed	Community-curated; 1800 taxa; variable annotation quality; unparalleled breadth
UTDAC2020 [24]	RGB stills	5168	Boxes (4 multiple species)	Expert	Balanced	5–20	Med–High	Competition-grade; refined annotations; 2024 adds murkier imagery and adversarial distractors
URPC [25]	RGB video	4.7 k frames	Boxes (4 multiple species)	Expert	Moderate imbalance	3–12	Med	Robot picking contest; scallops, echinus, starfish, holothurians; mobile platforms
MFT25 [26]	Video	48 k frames	Boxes + IDs	Expert	Balanced	1–8	Low–Med	408 k annotations; 15 clips; dense multi-target tracking with occlusions and erratic motion

As shown in Table 1, newer datasets like MFT25 and AquaDeep prioritise dense annotations and multimodality, reflecting the field's shift from simple benthic surveys to complex tracking and industrial

monitoring tasks. The progression from static RGB imagery (Brackish, URPC) to video sequences with temporal coherence (Fish4Knowledge, MFT25) enables research on multi-object tracking and behaviour analysis, while the integration of auxiliary modalities such as water-quality logs in AquaDeep or taxonomic metadata in FathomNet supports context-aware detection that goes beyond pixel-level features alone.

4.1. Dataset coverage trends

Classical coastal datasets (URPC, Brackish) focus on a handful of benthic species but capture photometric extremes such as tidal plumes and artificial lighting. Reef-focused corpora (Fish4Knowledge, DeepFish) provide longer sequences with schooling behaviour, enabling research on temporal coherence, while FathomNet extends taxonomic breadth with imagery from deep ROV missions. Challenge-driven releases like FishCLEF and Underwater Target Detection and Classification (UTDAC) introduce annual updates with hidden test sets, encouraging robust cross-site generalisation and transparent leaderboards.

Newer resources prioritise either dense annotations or multimodality. MFT25 deliberately annotates every individual across thousands of frames, supporting multi-object tracking research that goes beyond frame-level detection [26]. AquaDeep couples RGB feeds with hatchery water-quality metadata so that detectors can be co-trained with contextual signals, a prerequisite for fault diagnosis in commercial aquaculture [27]. SUIM, while originally released for segmentation, supplies high-quality masks that underpin physics-guided enhancement and synthetic data generation pipelines.

4.2. Benchmark protocols

Two competitions dominate evaluation practice. The URPC publishes a public training set and withholds test imagery, with submissions ranked by mAP@0.5 and latency under strict runtime budgets [25]. The UTDAC challenge alternates between still-image and video tracks and explicitly scores robustness under adversarial lighting or synthetic sediment bursts. FishCLEF-style benchmarks focus on species-level counting metrics, making direct cross-study comparisons unreliable [28].

4.3. Selection considerations

Researchers typically match datasets to the intended deployment domain: energy infrastructure inspections prefer URPC-like benthic imagery, conservation monitoring benefits from FathomNet's species breadth, and aquaculture adopts Brackish or AquaDeep because they mirror cage environments. Regardless of the choice, dataset bias remains a core risk [17]. Consequently, recent surveys advocate reporting not only image counts but also depth ranges, turbidity descriptors, annotation quality, and licensing constraints so that future studies can replicate or extend results without repeating exhaustive data collection.

4.4. Dataset quality and reproducibility challenges

While Table 1 highlights the diversity of underwater datasets, systematic quality issues and benchmark fragmentation hinder fair comparison and reproducibility. Annotation noise is pervasive: crowd-sourced labels can be inconsistent for morphologically similar taxa, and bounding box inconsistencies (tight *vs.* loose boxes) introduce training instability. Class imbalance severely affects datasets like Brackish, where

dominant species account for most annotations while rare benthic organisms have very few examples, forcing practitioners to apply aggressive oversampling or focal loss strategies that may not generalise.

Sensor heterogeneity compounds these challenges. DeepFish aggregates imagery from GoPros, Digital Single-Lens Reflex (DSLRs), and ROV-mounted cameras with varying resolutions (720p to 4K), white balance settings, and lens distortions, yet few studies report per-sensor performance breakdowns. Models trained on one sensor configuration often degrade when deployed on different hardware, a phenomenon rarely quantified in published benchmarks.

Benchmark fragmentation further complicates reproducibility. URPC reports mAP@0.5, UTDAC uses mAP@0.5:0.95, and FishCLEF emphasises species-level F1 scores, making cross-study comparisons unreliable. Evaluation scripts, data splits, and preprocessing pipelines are frequently undocumented or unavailable, forcing researchers to reverse-engineer baselines. Missing metadata such as exact camera intrinsics, lighting conditions, or water chemistry limits the ability to diagnose failure modes or adapt models to new sites.

Addressing these issues requires community-wide standards: publicly archived evaluation code, standardised train/val/test splits with fixed random seeds, and mandatory reporting of annotation protocols, sensor specifications, and environmental metadata. Until such practices become routine, the underwater detection literature will remain fragmented, and claimed performance gains may reflect dataset idiosyncrasies rather than genuine algorithmic advances.

5. Deep learning for underwater detection

Deep learning models have become the standard for underwater perception due to their ability to disentangle colour casts, clutter, and deformable shapes better than handcrafted pipelines [9,14]. However, architectural choices dictate whether detectors deliver sufficient accuracy, speed, and robustness for field deployment. This section synthesises the main model families and highlights trends observed in recent literature.

5.1. CNN-based detectors

Two-stage detectors such as Faster R-CNN and Mask R-CNN [29, 30] remain the backbone for high-precision surveys, especially when researchers require per-instance segmentation or uncertainty estimates. Underwater variants integrate attention modules or adversarial occlusion branches to handle camouflage, as seen in [31], which reports improved robustness on UTDAC by explicitly modelling self-occlusion. Single-stage CNNs gained popularity because they better satisfy embedded compute budgets. SSD [32] initiated this shift, but the YOLO family [33–35] ultimately became the default choice for aquaculture monitoring [36], offshore inspection, and AUV deployments, thanks to lightweight backbones and mature tooling.

5.2. YOLO and task-specific optimisation

YOLO derivatives perform best when paired with domain-aware feature fusion. Zhang *et al.* [37] augmented YOLOv4 with an attentional feature fusion module (AFFM), boosting URPC mAP from 84.9% to 92.7%. Pedersen *et al.* [38] demonstrated that careful anchor scaling and turbidity augmentations narrow the gap

between lightweight YOLO and heavier two-stage baselines on the Brackish dataset. Lightweight detector heads and careful calibration enable embedded aquaculture deployment [36].

5.3. Transformer-based detectors

Transformers provide global receptive fields that help distinguish overlapping fish or translucent jellyfish. Detection Transformer (DETR) [39] and its deformable or real-time variants [40,41] inspired several underwater studies that replace convolutional necks with multi-head attention. Real-Time DEtection TRansformer (RT-DETR) style models reduce handcrafted priors, allowing consistent performance when water colour shifts between deployments. However, transformers often incur higher latency and memory usage, pushing researchers to distil them into CNN surrogates or pair them with lightweight enhancement modules before inference.

5.4. Hybrid and multimodal approaches

Hybrid detectors combine CNN backbones with transformer decoders or fuse optical imagery with sonar, Light Detection and Ranging (LiDAR), or physics-based priors. Medical-imaging-inspired designs [42] illustrate how transformer tokens can encode contextual cues while CNN heads preserve high-frequency localisation. Multimodal systems align sonar depth cues with optical RGB to disambiguate occlusions [15], whereas physics-based enhancement pipelines (e.g., Contrast Limited Adaptive Histogram Equalization (CLAHE), white balance, Retinex) or learned underwater Generative Adversarial Networks (GANs) [9,43,44] are inserted upstream of detectors to normalise colour and contrast. Such cascades often improve minority-class recall at the cost of additional latency.

5.5. Self-supervised, few-shot, and domain adaptation

Data scarcity remains the largest barrier to generalisable UOD. Researchers therefore exploit unsupervised knowledge transfer [45], self-supervised pretext tasks (colourisation, jigsaw puzzles), or few-shot finetuning that prioritises species newly discovered at a site. CycleGAN- or WaterGAN-style translation [44,46] generates synthetic underwater scenes from terrestrial images, augmenting rare classes without additional dives. Domain adaptation studies increasingly report results on SUIM, UTDAC, and FathomNet splits to quantify generalisation across water types [14].

5.6. Systematic comparison across architectures

To provide a structured analysis, Table 2 presents a systematic cross-model comparison synthesising performance, computational cost, and deployment characteristics across CNN, YOLO, DETR, and hybrid families. This comparison consolidates metrics from representative studies to highlight fundamental trade-offs between accuracy, speed, memory footprint, and robustness under varying underwater conditions. Table 2 highlights the usual deployment trade-offs. YOLO-family detectors remain the most practical option for onboard AUV/ROV inference, where latency, power, and throughput dominate design constraints. Two-stage CNNs typically offer stronger localisation and higher precision, but their slower inference makes them better suited to offline survey and post-mission analysis. DETR-style transformers provide global context and can handle dense or occluded scenes well, yet their heavier

compute and memory demands often restrict embedded use. Hybrid CNN–Transformer designs sit between these extremes, trading added architectural complexity for improved context modelling and robustness. Multimodal fusion is the most reliable under severe turbidity, but it comes with integration overhead (calibration, synchronisation) and additional hardware and latency costs.

Table 2. Systematic comparison of deep learning architectures for underwater object detection.

Family	Representative models	Typical metric(s)	Strengths	Limitations	Compute	Best-suited scenarios	Notes (caveats)
One-stage CNN detectors	YOLO (v3–v8), SSD-style variants	mAP@0.5/AP50; FPS	High throughput; simpler deployment	Struggles with small objects/heavy turbidity without enhancement	Low–Med	Real-time AUV/ROV onboard detection	Reported metrics are highly sensitive to input resolution and augmentation
Two-stage CNN detectors	Faster R-CNN, Mask R-CNN (detector head)	AP/AP50; recall-oriented	Stronger localisation; better for cluttered scenes	Higher latency; heavier memory footprint	Med–High	Offline analysis; high-precision inspection	Speed numbers depend strongly on backbone + proposal settings
Transformer-based detectors	DETR family, Deformable DETR, RT-DETR variants	COCO-style AP (0.5:0.95)	Global context; fewer hand-crafted priors	Data-hungry; training stability; can be slower on edge	Med–High	Complex backgrounds; multi-object scenes	Metric choice varies across papers; compare like-for-like only
Hybrid CNN–Transformer	CNN backbone+ transformer encoder/decoder, Swin backbones	AP/mAP; robustness-focused	Improved robustness +context while keeping CNN inductive bias	Complexity; harder to tune; inconsistent reporting	Med–High	Variable visibility; mixed-scale targets	Often paper-specific ablations; beware cherry-picked settings
Multimodal/ fusion-based	Optical+sonar fusion, enhancement+ detector pipelines	UIQM/UCIQE+ AP (if detection included)	Works when optical visibility collapses; complementary sensing	Calibration+ synchronisation overhead; dataset scarcity	High	Low-visibility missions; long-range detection	Not all fusion papers report detection mAP; avoid mixing metrics

Note: This table is architectural/qualitative. Reported accuracy and speed are not directly comparable across papers unless dataset, metric definition, input size, and hardware are matched.

5.7. Comparison of representative techniques

Table 3 collates recent underwater-specific detectors. Reported numbers vary across datasets, yet the table highlights common trade-offs: attention-enhanced YOLO variants reach high mAP while transformers improve multi-object tracking (MOT) metrics, and hybrid or multimodal systems offer robustness at additional computational cost.

The results in Table 3 reveal a clear trade-off landscape: YOLO-based detectors consistently achieve real-time frame rates (30–62 FPS) with competitive mAP scores, making them the preferred choice for resource-constrained deployments on embedded platforms like Jetson TX2 or Xavier NX. Transformer-based approaches, exemplified by the SU-T tracker, excel in dense multi-object scenarios where global context is critical, though at the cost of reduced throughput and higher memory consumption. Hybrid and multimodal systems, such as optical-sonar fusion, offer the greatest robustness under extreme turbidity or occlusion but require careful sensor calibration and incur additional latency, positioning them as specialist tools rather than general-purpose solutions.

Collectively, the literature demonstrates that no single architecture dominates across all deployment regimes. Instead, researchers balance accuracy, computational budget, and environmental robustness by mixing architectural innovations with dataset curation, enhancement pipelines, and specialised evaluation metrics, motivating the deeper dives presented in the following sections.

Table 3. Representative underwater detection techniques from the literature.

Study	Key Contribution and Performance
Pedersen *et al.* [38]	YOLOv3 with custom anchors on Brackish dataset. mAP@0.5: 83.7%, 45 FPS on Jetson TX2. Domain-specific anchors + turbidity augmentation narrow gap to two-stage baselines.
Zhang *et al.* [37]	YOLOv4+AFFM on URPC2019. mAP@0.5: 92.7%, 62 FPS on RTX2080. Multi-scale attentional fusion mitigates colour casts and small-object misses.
Hu *et al.* [36]	Improved YOLOv4 on aquaculture feed dataset. F1: 95.1%, 30 FPS on GTX1080. Lightweight head plus confidence calibration enables on-farm pellet detection.
Zeng *et al.* [31]	Faster R-CNN + occlusion net on UTDAC2020. mAP@0.5: 90.3%, 12 FPS on V100. Adversarial occlusion branch boosts robustness to overlapping holothurians.
Li *et al.* [26]	SU-T (Transformer+Unscented Kalman Filter (UKF)) on MFT25. Higher Order Tracking Accuracy (HOTA): 34.1, IDF1: 44.6, 18 FPS on A100. Joint detection-tracking pipeline tailored for dense multi-fish scenes.
Kim *et al.* [15]	Optical+Sonar fusion on custom ROV dataset. Underwater Image Quality Measure (UIQM) ↑ 94%, 8 FPS on Xavier NX. Physics-aware fusion improves detection in heavy turbidity where RGB alone fails.

5.8. Architectural trade-offs and design principles

Selecting an architecture requires balancing competing objectives: accuracy, inference speed, memory footprint, and robustness to domain shift. CNN *vs.* Transformer trade-offs center on receptive field versus computational cost. CNNs excel at capturing local textures and edges through hierarchical convolutions, making them efficient for real-time embedded deployment. However, their limited receptive fields struggle with global context in cluttered scenes where overlapping fish or translucent jellyfish require long-range dependencies. Transformers address this via self-attention mechanisms that model global interactions; deformable attention and iterative box refinement can improve convergence and performance in crowded scenes [41]. However, this incurs a significant cost: training Deformable DETR typically requires high-memory GPUs (e.g., $\geq$ 15 GB Video Random Access Memory (VRAM) in NVIDIA's Train, Adapt, and Optimize (TAO) reference) [47], and transformers may require longer training schedules to reach comparable accuracy.Two-stage vs one-stage detectors represent a precision-speed trade-off. Two-stage methods (Faster R-CNN, Mask R-CNN) generate region proposals before classification, achieving strong precision and enabling instance segmentation. This precision supports ecological surveys requiring per-individual counts and morphological measurements. However, their sequential architecture can be challenging to deploy under tight real-time constraints on embedded hardware. One-stage detectors (YOLO, SSD) predict bounding boxes directly from feature maps, typically offering higher throughput.

This trade-off is acceptable for applications where missing a few detections is tolerable but latency violations are catastrophic, such as collision avoidance or live fish counting in aquaculture.

Attention mechanisms improve performance when applied judiciously but introduce overhead. Attention modules (e.g., Convolutional Block Attention Module (CBAM), Squeeze-and-Excitation Network (SE-Net)) can suppress background clutter and emphasise foreground organisms, but add latency and memory usage that may exceed embedded platform budgets. Channel attention tends to be cheaper but less effective in highly cluttered scenes. Designers should apply attention selectively: at deeper layers where semantic features dominate, rather than at shallow layers where computational cost outweighs benefits.

Backbone depth yields diminishing returns. Increasing backbone depth typically yields modest improvements on underwater datasets, far below gains observed on terrestrial benchmarks like COCO. This is because underwater images suffer from low signal-to-noise ratios due to scattering and absorption, limiting the utility of very deep feature hierarchies. Practitioners often find that mid-depth backbones strike the best accuracy-efficiency balance, with deeper networks offering marginal gains at substantial computational cost.

5.9. Failure modes and limitations

Despite progress, underwater detectors suffer from systematic failure modes that limit real-world deployment. Turbidity-induced failures are frequent. High turbidity reduces contrast and introduces backscatter haze, causing detectors to miss low-contrast organisms (flatfish, camouflaged crustaceans) or hallucinate false positives from suspended particulates. Studies report substantial performance loss under domain shift, with mAP decreases on the order of 10–15 percentage points in controlled evaluations [18]. Color cast shifts where blue-green dominance in shallow water transitions to monochromatic gray in deep or murky conditions further confuse RGB-trained models that rely on color cues for species discrimination.

Domain shift challenges manifest when training and deployment environments differ. Models trained on fixed-rig datasets (Brackish) perform well on in-distribution test sets but degrade when evaluated on mobile platform data (URPC, Fish4Knowledge) due to motion blur, varying camera angles, and dynamic lighting. Geographic domain shift is equally severe: detectors trained on Australian coral reefs (DeepFish) suffer substantial mAP loss when applied to North Atlantic kelp forests or Arctic waters, where species morphology, background textures, and illumination conditions differ dramatically.

Interactions between enhancement and detection pipelines create subtle failure modes. Over-enhancement applying aggressive histogram equalization or color correction can amplify noise, create artificial edges, and saturate regions, leading detectors to produce spurious bounding boxes around enhancement artifacts. Conversely, under-enhancement leaves images degraded, causing detectors to miss small or low-contrast targets. The optimal enhancement strategy is dataset- and model-dependent: YOLO models benefit from moderate contrast enhancement (e.g., CLAHE with conservative clip limits), while transformers are more robust to raw imagery due to their global context modeling. Few studies systematically ablate enhancement pipelines, making it difficult to disentangle photometric preprocessing gains from architectural improvements.

Small object detection remains difficult in underwater scenarios where juvenile fish, shrimp, or distant organisms occupy only a few dozen pixels. Standard anchor-based detectors struggle because default anchor scales (designed for terrestrial objects like cars and pedestrians) mismatch underwater targets. Specialized small-object detectors (e.g., multi-scale feature pyramids, attention-guided refinement) can improve recall but at the cost of increased false positives from marine snow or debris. Translucent and camouflaged species jellyfish, cuttlefish, flatfish remain challenging because their low texture and color similarity to backgrounds defeat CNN edge detectors.Transformer models show modest improvements by leveraging global context, but these species still exhibit substantially lower recall than opaque, high-contrast targets like starfish or scallops.

5.10. Domain adaptation bottlenecks

Cross-dataset generalization is a significant bottleneck. Performance drops are quantifiable and severe: models achieving good mAP on one dataset often degrade substantially on another, despite both featuring similar target classes. This gap persists even after standard data augmentation (flips, crops, color jitter), indicating that photometric and geometric augmentations alone cannot bridge domain gaps caused by sensor differences, depth variations, and ecological context shifts.

Transfer learning limitations are evident. ImageNet pretraining ubiquitous in terrestrial detection provides only limited mAP gains on underwater datasets, far below the substantial improvements observed on COCO or Pascal VOC. This is because ImageNet's terrestrial images (animals, vehicles, indoor scenes) share little visual similarity with underwater environments (blue-green color casts, suspended particulates, deformable organisms). Self-supervised pretraining on large unlabeled underwater corpora (e.g., Fish4Knowledge [22]) shows promise, yielding notable mAP improvements, but requires substantial computational resources and remains underexplored.

Unsupervised domain adaptation techniques such as adversarial feature alignment (Domain-Adversarial Neural Network (DANN), Adversarial Discriminative Domain Adaptation (ADDA)) or CycleGAN-based image translation can reduce domain gaps but introduce new failure modes. Adversarial alignment can over-smooth features, losing fine-grained species distinctions, while CycleGAN translation sometimes generates unrealistic artifacts (e.g., synthetic fish textures that don't match real morphology), degrading detector performance Few-shot learning constraints are equally limiting: achieving strong performance with only a handful of examples per class (common for rare species) requires sophisticated meta-learning or prototypical networks, yet these methods remain brittle and sensitive to the choice of support set, often failing to generalize beyond the specific few-shot scenarios they were trained on.

6. Enhancement and detection pipelines

Underwater perception systems rarely deploy monolithic detectors. Instead, practitioners assemble pipelines that compensate for optical degradation, balance compute budgets, and enforce temporal consistency before final detections reach decision logic. Three themes dominate recent literature.

6.1. Classical enhancement + CNN detection

Many fielded systems retain differentiable versions of classical pre-processing—white balance, homomorphic filtering, CLAHE, and Retinex—to restore contrast prior to detection [9]. These operations are lightweight enough for ARM-based System-on-Chips (SoCs) and stabilise the colour statistics that YOLO or Faster R-CNN expect. Hu *et al.* [36] combined histogram equalisation with a pruned YOLOv4 head to achieve reliable aquaculture feed monitoring on embedded GPUs, illustrating how modest enhancement can prevent false alarms without redesigning the detector.

6.2. GAN- and physics-based restoration

When water colour varies widely, data-driven restoration is preferred. Conditional GANs such as Underwater-GAN [43] or translation models like CycleGAN/WaterGAN [44, 46] synthesise realistic turbidity patterns and learn inverse mappings to recover latent colours. Recent survey papers report that chaining GAN restoration with detector fine-tuning can improve minority-class recall modestly on Brackish and URPC splits, albeit with extra latency. Physics-aware networks that embed attenuation priors directly into the generator reduce artefacts and maintain linear colour spaces for downstream measurements.

6.3. Joint enhancement-detection and multimodal fusion

An emerging trend is end-to-end optimisation where enhancement, detection, and even tracking share gradients. Recent work explores integrated pipelines that accept contrast-normalised features from enhancement modules, while transformer decoders model long-range dependencies across time. Multimodal pipelines fuse sonar depth cues with optical imagery [15], enabling robust detections when turbidity saturates the camera. Others exploit temporal filtering or lightweight Kalman smoothers to enforce track continuity, particularly in MultiFish/MFT25-style scenarios [26]. These integrated pipelines demand more computation but deliver resilience to lighting changes, motion blur, and occlusion, which is indispensable for autonomous AUV or ROV deployments where mission aborts are costly.

Overall, the literature underscores that enhancement choices interact strongly with detector architecture and hardware budgets. Pipelines must therefore be co-designed with deployment constraints such as VRAM limits, strobed lighting artefacts, and uplink bandwidth, rather than appended as afterthoughts.

7. Evaluation metrics and benchmarking

Evaluating underwater detection methods requires interpreting results within the context of evaluation practices. In contrast to terrestrial detection, underwater deployments prioritize photometric fidelity, temporal stability, and resource usage. This section collates the metrics and benchmarking conventions that appear across URPC, FishCLEF, UTDAC, and recent journal articles.

7.1. Detection metrics

Most studies follow COCO or PASCAL Visual Object Classes (VOC) conventions, reporting precision, recall, and mAP over IoU thresholds [48,49]. mAP@0.5 remains the primary metric in URPC, yet stricter

ranges (mAP@0.5:0.95) are increasingly emphasised because bounding boxes must respect fine-grained morphologies (e.g., starfish arms). Researchers also report class-wise recall to reflect ecological priorities missing a protected species may be costlier than occasional false alarms. Precision-recall curves remain important for tuning threshold-dependent detectors when visibility suddenly degrades.

7.2. Beyond mAP: marine-specific KPIs

Marine robotics groups augment mAP with metrics that capture operational impact. URPC introduces latency caps and counts frames processed per second to ensure real-time grasp planning, while aquaculture deployments report per-frame fish counting accuracy and overfeeding alarms [36]. Image-quality indices such as UIQM and Underwater Color Image Quality Evaluation (UCIQE) quantify how enhancement pipelines alter colour balance [50,51], helping practitioners separate photometric gains from detector improvements. For video-centric tasks, tracking metrics like CLEAR MOT, HOTA, and IDF1 [26,52] evaluate temporal consistency, penalising ID switches that could corrupt population estimates.

7.3. Benchmark protocols

URPC publishes labelled training data but withholds test imagery; submissions include predictions and runtime logs that organisers evaluate centrally. UTDAC follows a similar structure yet injects synthetic sediment bursts and lighting attacks to stress-test robustness, ranking teams on combined accuracy and stability scores. FishCLEF focuses on species-level F1 and requires per-clip predictions, rewarding systems that maintain identity across short video bursts. Researchers often complement these competitions with cross-dataset tests, training on Brackish or URPC and evaluating on DeepFish or FathomNet to expose dataset bias and domain shift [14,17].

7.4. Failure analysis and stress testing

Quantitative summaries are increasingly paired with qualitative diagnostics. Studies report confusion matrices that highlight translucent or deformable classes (jellyfish, shrimp) as persistent failure cases [37]. Robustness evaluations perturb imagery with synthetic turbidity, blur, or colour jitter to estimate degradation rates, an approach motivated by field observations where weather fronts quickly change visibility [14]. Some teams log power consumption, VRAM usage, and on-board latency to capture the energy budget and thermal constraints of AUV deployments [36]. These peripheral metrics remain unevenly reported but are critical for translating leaderboard gains into deployable systems.

The benchmarking ecosystem indicates a shift from single-number mAP reporting to richer scorecards that capture energy, latency, temporal stability, and species-specific risk. Future work should focus on harmonising these metrics so that new datasets and architectures can be compared without reproducing each bespoke evaluation script.

8. Open challenges and future directions

Despite steady progress, several obstacles prevent underwater object detection from matching terrestrial robustness.

8.1. *Persistent challenges*

- Long-tail species and data scarcity. Most datasets remain skewed toward crabs, starfish, or large fish, leaving fragile species underrepresented. Few-shot learning and synthetic augmentation (CycleGAN, WaterGAN) help but still require careful validation to avoid artefacts.
- Dynamic water properties. Turbidity, illumination, and particulates can change faster than models can be retrained. Continual learning on edge devices, coupled with uncertainty monitoring, is needed to adapt without catastrophic forgetting.
- Hardware and energy limits. Battery-powered AUVs must balance perception quality with propulsion demands. Profiling latency, VRAM, and power draw as advocated by [36] should become standard reporting practice so algorithms can be matched to hardware tiers.
- Benchmark fragmentation. URPC, UTDAC, FishCLEF, and proprietary datasets all use different splits and metrics, complicating meta-analysis. Community benchmarks that release harmonised protocols will accelerate reproducibility.

8.2. *Emerging research directions*

- Vision-language and foundation models. Multimodal encoders inspired by CLIP [53] can jointly embed textual species descriptions and visual cues, aiding rare-species recognition and semi-automated annotation. SAM-like segmentation foundation models may reduce annotation cost and support segmentation-driven pipelines in underwater scenes [54]. Vision-language pretraining on paired dive logs and imagery is beginning to appear in marine biodiversity studies and could drastically reduce labelling cost.
- Multimodal fusion and 3D perception. Combining sonar cues with optical imagery can stabilise detections during turbidity spikes [15]. Structure-from-motion reconstructions or acoustic tomography enable coarse 3D localisation, which helps disambiguate overlapping organisms. Few-shot, self-supervised, and domain-adaptive learning. Unsupervised knowledge transfer [45] and adaptive augmentation [14] should be extended to transformer-based detectors, while new datasets such as MFT25 and AquaDeep encourage research on temporal and contextual cues.
- Onboard deployment and responsive autonomy. Embedded-friendly transformers, sparsely updated detectors, and hardware-aware neural architecture search are critical for real-time decisions on AUVs, ROVs, and low-cost edge devices.

8.3. *Responsible and sustainable AI*

Ethical deployment demands transparency about dataset provenance, sensor footprints, and potential misuse. Bias toward specific regions or species can skew conservation decisions; hence authors should document collection methods, licensing, and socio-ecological context [55]. Sustainability considerations include minimising dive time, sharing trained weights to avoid redundant missions, and aligning detection goals with community stakeholders (fisheries, regulators, indigenous groups) to ensure data sovereignty [56]. Emerging regulatory frameworks are likely to mandate audit trails for marine monitoring AI, making reproducible reporting as important as algorithmic novelty.

9. Conclusion

This survey reviewed underwater imaging challenges, datasets, model architectures, enhancement pipelines, and evaluation practices. Progress stems from co-designing detectors with domain-aware restorations, diversified datasets, and metrics that reflect ecological stakes rather than headline mAP alone. CNN-based YOLO variants remain attractive for embedded deployments, transformers excel at modelling dense multi-object scenes, and hybrid or multimodal systems provide robustness when turbidity or occlusion dominate.

Remaining gaps include long-tail species coverage, harmonised benchmarks, and lightweight adaptation techniques that can track water-quality shifts without exhaustive retraining. Closing these gaps will take coordinated effort among field operators, dataset curators, and algorithm designers. Sharing open protocols, reporting latency and energy alongside accuracy, and investing in responsible AI practices will transform underwater object detection from bespoke deployments into repeatable, trustworthy infrastructure for marine science, aquaculture, and environmental stewardship.

Data availability statement

No supplementary or additional data were generated in this study.

Declaration of generative AI and AI-assisted technologies

The authors declare that no generative AI or AI-assisted technologies were used in the writing, data analysis, or preparation of this manuscript.

Acknowledgments

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/Y000773/1.

Authors' contribution

Conceptualization, P.L. and H.B.; methodology, P.L. and H.B.; software, H.B.; validation, P.L. and H.B.; formal analysis, H.B.; investigation, H.B.; resources, P.L.; data curation, H.B.; writing—original draft preparation, H.B.; writing—review and editing, P.L. and H.B.; visualization, H.B.; supervision, P.L.; project administration, P.L.; funding acquisition, P.L. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Pengcheng Liu holds the position of Associate Editor for *Robot Learning* and has not peer reviewed or made any editorial decisions for this paper.

References

[1] Copley J. Just how little do we know about the ocean floor? 2014. Available: https://www.scientificamerican.com/article/just-how-little-do-we-know-about-the-ocean-floor/ (accessed on 24 January 2026).

[2] US Geological Survey. How much water is there on earth? 2016. Available: https://www.usgs.gov/water-science-school/science/how-much-water-there-earth (accessed on 24 January 2026).

[3] Craig RK. Marine biodiversity, climate change, and governance of the oceans. *Diversity* 2012, 4(2):224–238.

[4] Doney SC, Ruckelshaus M, Duffy JE, Barry JP, Chan F, *et al.* Climate change impacts on marine ecosystems. *Annu. Rev. Mar. Sci.* 2012, 4:11–37.

[5] Tashim NAZ, Lim TH, Zariful MW, Liu P. Regression-based artificial intelligence length and weight estimation for sustainable prawn aquaculture. *Smart Agric. Technol.* 2025, 12:101089.

[6] Zariful W, Tashim NAZ, Lim TH, Basri AM, Chuprat S, *et al.* Analysis of deep learning algorithms for prawn aquaculture in a challenging environment. In *Proceedings of 2023 6th International Conference on Applied Computational Intelligence in Information Systems (ACIIS)*, Bandar Seri Begawan, Brunei Darussalam, October 23–25, 2023, pp. 1–5.

[7] Abd Zariful MW, Tashim NAZ, Lim TH, Fakhrurroja H, Chuprat S, *et al.* Comparison of biocode based machine learning and segmentation model for automated prawn size prediction for real prawn farm. In *Proceedings of 2024 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Dalian, China, April 12–14, 2024, pp. 581–586.

[8] Yang S, Liu P, Lim TH. IoT-based underwater robotics for water quality monitoring in aquaculture: a survey. In *Proceedings of 2023 International Conference on Robot Intelligence Technology and Applications*, Taicang, China, December 6–8, 2023, pp. 32–42.

[9] Li C, Guo J, Cong R, Pang Y, Wang B. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *IEEE Trans. Image Process.* 2016, 25(12):5664–5677.

[10] Mobley C. *Light and Water: Radiative Transfer in Natural Waters*. New York: Academic Press, 1994.

[11] Sun L, Zhao C, Yan Z, Liu P, Duckett T, *et al.* A novel weakly-supervised approach for RGB-D-based nuclear waste object detection. *IEEE Sens. J.* 2018, 19(9):3487–3500.

[12] Liu P, Huda MN, Sun L, Yu H. A survey on underactuated robotic systems: bio-inspiration, trajectory planning and control. *Mechatronics* 2020, 72:102443.

[13] Zhang B, Liu P. Model-based and model-free robot control: a review. In *RiTA 2020: Proceedings of the 8th International Conference on Robot Intelligence Technology and Applications*, Cardiff, UK, December 11–13, 2020, pp. 45–55.

[14] Yuan X, Guo L, Luo C, Zhou X, Yu C. A survey of target detection and recognition methods in underwater turbid areas. *Appl. Sci.* 2022, 12(10):4898.

[15] Kim HG, Seo J, Kim SM. Underwater optical-sonar image fusion systems. *Sensors* 2022, 22(21):8445.

[16] Liu P, Yu H, Cang S. Adaptive neural network tracking control for underactuated systems with matched and mismatched disturbances. *Nonlinear Dyn.* 2019, 98(2):1447–1464.

[17] Torralba A, Efros AA. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, USA, June 20–25, 2011, pp. 1521–1528.

[18] Folkman L, Pitt KA, Stantic B. A data-centric framework for assessing performance of object detection algorithms in underwater environments. *Appl. Intell.* 2025, 55(4):272.

[19] Pedersen M, Haurum JB, Gade R, Moeslund TB. Detection of marine animals in a new underwater dataset with varying visibility. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, USA, June 16–20, 2019, pp. 18–26.

[20] Islam MJ, Edge C, Xiao Y, Luo P, Mehtaz M, *et al.* Semantic segmentation of underwater imagery: dataset and benchmark. *IEEE Rob. Autom. Lett.* 2020, 5(2):3037–3044.

[21] Saleh A, Laradji IH, Konovalov DA, Bradley M, Vazquez D, *et al.* A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Sci. Rep.* 2020, 10(1):14671.

[22] Fish4Knowledge Project. 2010. Available: https://homepages.inf.ed.ac.uk/rbf/fish4knowledge/ (accessed on 24 January 2026).

[23] Katija K, Orenstein E, Schlining B, Lundsten L, Barnard K, *et al.* FathomNet: a global image database for enabling artificial intelligence in the ocean. *Sci. Rep.* 2022, 12(1):15914.

[24] FlyAI. Underwater target detection algorithm competition (UTDAC2020) dataset page. 2020. Available: https://universe.roboflow.com/utdac2020/utdac2020-iqq77/dataset/1 (accessed on 24 January 2026).

[25] URPC Organizing Committee. Underwater robot picking contest (URPC). 2021. Available: http://2021en.urpc.org.cn/a/dhyw/ (accessed on 24 January 2026).

[26] Li W, Liu Y, Guo Q, Wei Y, Leo HL, *et al.* When trackers date fish: a benchmark and framework for underwater multiple fish tracking. *arXiv* 2025, arXiv:2507.06400.

[27] AquaDeep Inc. AquaDeep—AI solutions for sustainable aquaculture. 2024. Available: https://www.aquadeep.ai/ (accessed on 24 January 2026).

[28] Joly A. FishCLEF dataset. 2025. Available: https://zenodo.org/records/15202605 (accessed on 24 January 2026).

[29] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, December 7–12, 2015, pp. 91–99.

[30] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 22–29, 2017, pp. 2980–2988.

[31] Zeng L, Sun B, Zhu D. Underwater target detection based on faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* 2021, 100:104190.

[32] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, *et al.* SSD: single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, October 11–14, 2016, pp. 21–37.

[33] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, June 27–30, 2016, pp. 779–788.

[34] Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* 2022, arXiv:2207.02696.

[35] Ultralytics Team. YOLOv8: new state-of-the-art detector (version 8). 2023. Available: https://github.com/ultralytics/ultralytics (accessed on 24 January 2026).

[36] Hu X, Liu Y, Zhao Z, Liu J, Yang X, *et al.* Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network. *Comput. Electron. Agric.* 2021, 185:106135.

[37] Zhang M, Xu S, Song W, He Q, Wei Q. Lightweight underwater object detection based on YOLOv4 and multi-scale attentional feature fusion. *Remote Sens.* 2021, 13(22):4706.

[38] Pedersen M, Haurum JB, Gade R, Moeslund TB. Detection of marine animals in a new underwater dataset with varying visibility. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, USA, June 16–20, 2019, pp. 18–26.

[39] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, *et al.* End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, online, August 23–28, 2020, pp. 213–229.

[40] Zhu X, Su W, Lu L, Li B, Wang X, *et al.* Deformable DETR: deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 3–7, 2021.

[41] Zhao Y, Lv W, Xu S, Wei J, Wang G, *et al.* RT-DETR: DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, June 16–22, 2024.

[42] Küçük DB, Imak A, Özçelik STA, Çelebi A, Türkoğlu M, *et al.* Hybrid CNN-transformer model for accurate impacted tooth detection in panoramic radiographs. *Diagnostics* 2025, 15(3):244.

[43] Yu X, Yang S, Qu Y, Hong M. Underwater-GAN: underwater image restoration via conditional generative adversarial network. In *Proceedings of the 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, August 20–24, 2018, pp. 66–75.

[44] Li J, Skinner KJ, Eustice RM, Johnson-Roberson M. WaterGAN: unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Rob. Autom. Lett.* 2018, 3(1):387–394.

[45] Zurowietz M, Nattkemper TW. Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration. *IEEE Access* 2020, 8:143558–143568.

[46] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 22–29, 2017, pp. 2223–2232.

[47] NVIDIA TAO Toolkit v5.3.0 documentation: deformable DETR. 2024. Available: https://docs.nvidia.com/tao/archive/5.3.0/text/object_detection/deformable_detr.html (accessed on 24 January 2026).

[48] Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision* 2010, 88(2):303–338.

[49] Lin TY, Maire M, Belongie S, Hays J, Perona P, *et al.* Microsoft COCO: common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, September 6–12, 2014, pp. 740–755.

[50] Panetta K, Gao C, Agaian S. Human-visual-system-inspired underwater image quality measures. *IEEE J. Oceanic Eng.* 2016, 41(3):541–551.

[51] Yang M, Sowmya A. An underwater color image quality evaluation metric. *IEEE Trans. Image Process.* 2015, 24(12):6062–6071.

[52] Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP J. Image Video Process.* 2008, 2008(1):1–10.

[53] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, online, July 18–24, 2021, pp. 8748–8763.

[54] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, October 2–3, 2023.

[55] Hrckova A, Renoux J, Tolosana-Calasanz R, Chuda D, Tamajka M, *et al.* AI research is not magic, it has to be reproducible and responsible: challenges in the AI field from the perspective of its PhD students. *arXiv* 2024, arXiv:2408.06847.

[56] Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, *et al.* The CARE principles for indigenous data governance. *Data Sci. J.* 2020, 19(43):1–12.