A Preliminary Study on Underwater Transparent Objects Detection with Stereo Vision: Air vs Freshwater

Aaron Smiles[1], Changjae Oh[1], and Ildar Farkhatdinov[2]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, UK, a.l.smiles@qmul.ac.uk
[2] School of Biomedical Engineering and Imaging Sciences, Kings College London, UK

Abstract. This paper investigates how underwater conditions affect the detection and distance estimation of transparent plastic bottles using a stereo computer vision pipeline. We created a multi-environment dataset of stereo recordings with per-frame range logs in air and freshwater across common bottle types (empty and filled, smooth and textured) and marked distances. Compared with air, fresh water environment consistently increased errors and reduced reliability: the root mean squared distance error for the filled transparent bottle rose from 38 mm to 696 mm, and detection accuracy for textured empty bottles fell from 100% to 74.4%. A simple image enhancement increased the share of valid object-distance estimates from 41.6% to 50.8% but also raised the root mean squared distance error from 146.2 mm to 209.3 mm, revealing a trade-off between stability and precision. Selecting more stable viewing regions and filtering unreliable detections improved robustness (for example, reducing object-level error for the filled transparent bottle in air from 445 mm to 13 mm), but closing the gap between air and water will require methods that explicitly account for refraction and underwater imaging effects.

Keywords: underwater computer vision · stereo vision · transparent objects · refractive media · detection and ranging

1 Introduction

Robust perception of everyday *transparent plastic bottles* underwater is a persistent challenge for field robotics. This paper scopes a *stereo vision detection–and–ranging (D&R)* baseline—combining learned 2D detection with 3D range from a calibrated stereo rig—and systematically quantifies how moving from air to freshwater alters detection reliability and range accuracy across representative bottle types and viewpoints. The problem is difficult because refractive index mismatch, turbidity, low contrast and backscatter violate single-medium assumptions in commodity stereo pipelines, yielding invalid or biased depths (including not-a-number, "NaN", returns) precisely for filled transparent targets. Understanding the magnitude and structure of these errors matters for pollution monitoring, retrieval and manipulation, and telepresence safety: a data-centric

baseline that exposes failure modes and simple robustness gains is a prerequisite for progressing to geometry/refraction-aware methods and more reliable underwater systems.

Related work. In-air transparent-object perception has advanced via geometry-aware depth completion and 3D shape recovery (ClearGrasp [1]), geometry-guided monocular 6D pose estimation (GDR-Net) and its edge-fusion variant [2,3], and transformer- or reflection-based segmentation methods (Trans2Seg; glass segmentation with reflection priors [4,5]). Multiview perception (MVTrans [6]) and augmentation strategies such as FakeMix [7] further extend the detection of transparent objects. These approaches largely target pose or segmentation in clear, in-air environments, while we study *underwater* stereo detection-and-ranging of filled transparent bottles, where refraction, backscatter, and not-identitfied estimates (reported as not-a-number) or rear-wall depth estimates dominate the failure modes. These approaches largely target pose or segmentation in clear, in-air environments, while we study *underwater* stereo detection-and-ranging of filled transparent bottles, where refraction, backscatter, and not-identified estimates (reported as not-a-number) or rear-wall depth estimates dominate the failure modes.

This research extends our **previous works** [8,9], making an order of magnitude more data point observations, automating point distance, applying a basic enhancement inspired by related works, and recording stereo vision frame data for reproducibility and future enhancements. Specifically, our earlier simulator paper [8] implemented a calibrated ZED-based stereo vision stack for a mixed-reality teleoperation simulator, establishing the real-time detection–ranging pipeline, logging, and evaluation scaffolding we build upon here. Our subsequent study [9] quantified how moving from air to freshwater degrades bottle detection and stereo ranging—especially for water-filled transparent bottles—and showed that a lightweight enhancement can increase valid range coverage while slightly increasing error, motivating the multi-environment analysis we preliminarily report in this paper.

2 Method

The experiment aimed to quantify how moving from *Air* to *Freshwater* conditions alters transparent-object detection and stereo range estimation across representative bottle types and viewing positions. Throughout, we use "Air" and "Freshwater" (capitalised) as labels for the two environments; "Freshwater" denotes tap water in a windowed tank. We measured two range estimators — the manual Average Point Distance and the ZED SDK Object Distance — together with classification accuracy and detector confidence, using vertical sampling (Upper, Centre, Lower) to probe view dependence. We expected refractive index mismatch and contrast loss underwater to increase range error and reduce accuracy/confidence, with the largest degradations for filled transparent bottles; at the same time, we sought to establish a baseline and gain a clearer understanding of how such cases might eventually be mitigated. Finally, we examined whether simple pre-processing and

confidence-gated "success-only" filtering could recover partial robustness without retraining the detector.

System and data. A calibrated Stereolabs ZED Mini stereo camera observes bottle targets at marked distances in a windowed tank rig (Fig. 1). Inputs are live stereo or Stereolabs SVO2 stereo-video recordings; depth is computed by the Stereolabs ZED software development kit (SDK) and reported in millimetres [11]. Targets include smooth and textured bottles, filled variants, and a small medicine bottle across the Air and Freshwater conditions.

Detection and measurements. A YOLOv8 detector [10] processes the left image. The detected bounding boxes are then provided to the ZED SDK as detection-box objects via its custom bounding-box interface (type `CustomBoxObjectData`), so the SDK returns per-object 3D positions, dimensions, persistent IDs, and tracking. For each accepted detection we (i) record the SDK's per-detection *Object Distance* (Z range at the detection) and (ii) sample stereo range at three vertical regions of interest (ROI): *Upper*, *Centre*, *Lower* (offsets ± 50 px). We log per-frame: ground-truth range, environment and bottle-type labels, the detector's predicted class and confidence, and whether a valid range was returned by the SDK.

Metrics. We evaluate two representative conditions—*Air* and *Freshwater*. We report the six metrics used in the figures: (i) RMSE for *Avg Point Distance* (the average of Upper/Centre/Lower samples), (ii) RMSE for *Object Distance (All)* across all detections that return a numeric value, including zeros (non–not-a-number (NaN) samples), (iii) RMSE for *Object Distance (Success Only)* restricted to non-zero detections, (iv) *Classification Accuracy* (the share of detections where the model correctly said "`bottle`". Right/wrong only; confidence doesn't matter), (v) *Mean Confidence* (average confidence score the model assigned to its detections, whether they were right or wrong.), and (vi) *Mean Confidence (Success Only)* (average confidence restricted to correctly classified `bottle` detections). Aggregations are by bottle type and environment, matching the panel organisation in the figures.

Minimal ablation. We evaluate a light, inference-time pre-processing variant (*basicEnhance*) on the Freshwater subset; detector weights are unchanged. In this paper, *basicEnhance* refers to a simple in-house pre-processing technique that increases local contrast and edge strength; it is applied only at inference time and is not a third-party library. The same six metrics and aggregations are used to quantify any changes relative to the baseline.

3 Results

We report root mean squared error (RMSE), classification accuracy, and detector confidence across Air and Freshwater. We use *Average Point Distance* for the manual range estimate (average of the three vertical samples: Upper, Centre, Lower) and *Object Distance* for the ZED SDK's per-object range. Where noted, "success-only" excludes invalid ranges (e.g., NaN).

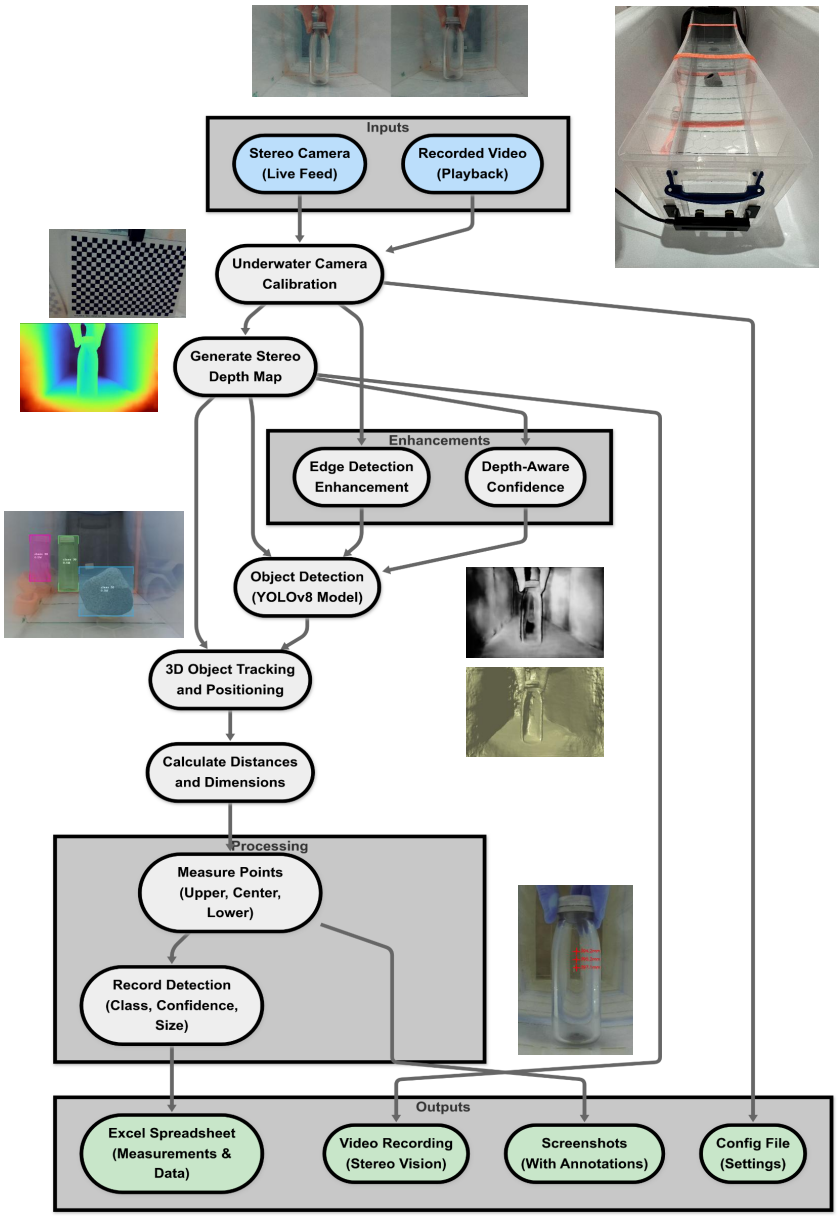

Fig. 1: End-to-end processing pipeline: stereo calibration/depth → optional edge/depth-aware confidence → YOLOv8 detection → ZED SDK ingestion/tracking → per-detection measurements and logs.

3.1 Cross-environment summary and manual vs automatic range.

Across bottle types, Freshwater environment increases error and reduces classifier performance relative to Air condition (Fig. 2). For point-wise range, RMSE rises for plastics, with the largest jump for the filled smooth (transparent) bottle (38→696 mm) and smaller increases for smooth/textured empty (Fig. 2a). An exception is the medicine bottle, which improves underwater (63→48 mm). For object-level range, RMSE generally increases when using all detections (Fig. 2b), with one exception (filled smooth: 445→432 mm). Restricting to success-only makes the trend unambiguous—errors rise underwater for every type and most strongly for the filled smooth bottle (13→333 mm) and textured empty (31→154 mm) (Fig. 2c). Classifier accuracy and mean confidence also drop underwater (e.g., textured empty accuracy 100→74.4%; confidence 77.4→59.1%), while medicine and textured filled remain highly accurate despite lower confidence (Fig. 2d-f). Manual Average Point Distance and automatic Object Distance differ in absolute scale, but they agree on relative difficulty and rank bottle types the same way: filled transparent is consistently the hardest underwater, while medicine is comparatively robust. The automatic Object Distance typically yields equal or larger RMSE than the manual measure under the same condition, reflecting depth-noise accumulation over the detection ROI and detector/tracker variability. Success-only filtering improves absolute figures yet leaves a clear gap between Air and Freshwater (Fig. 2(c,f)).

3.2 Point Distance.

Fig. 2(a) compares RMSE of per-pixel point distance across bottle types. Relative to Air, Freshwater generally increases error for plastic bottles, with the largest degradation for the filled smooth (transparent) bottle (38→696 mm) and moderate increases for smooth empty (34→127 mm) and textured empty (25→96 mm). In contrast, the medicine bottle improves underwater (63→48 mm).

Vertical sampling. Fig. 3 shows the samples which are tightly clustered in Air (top plots), but spread out underwater (bottom plots), especially for the filled smooth bottle, with occasional high-error spikes (failures). Textured empty remains comparatively stable across vertical positions.

Takeaway. Underwater conditions chiefly hurt point-wise accuracy for transparent plastics—dramatically when filled—while textured surfaces and the medicine bottle are more robust.

3.3 Object Distance.

Fig. 2(b) (all detections) shows that Freshwater generally increases object-level RMSE versus Air for most bottle types: medicine (437→508 mm), smooth empty (450→512 mm), textured empty (444→518 mm), textured filled (362→394 mm). The filled smooth (transparent) bottle is the only exception in (b), showing a small decrease (445→432 mm). Fig. 2c (success-only detections) reduces absolute errors for both media but exposes a clear underwater gap across all types:

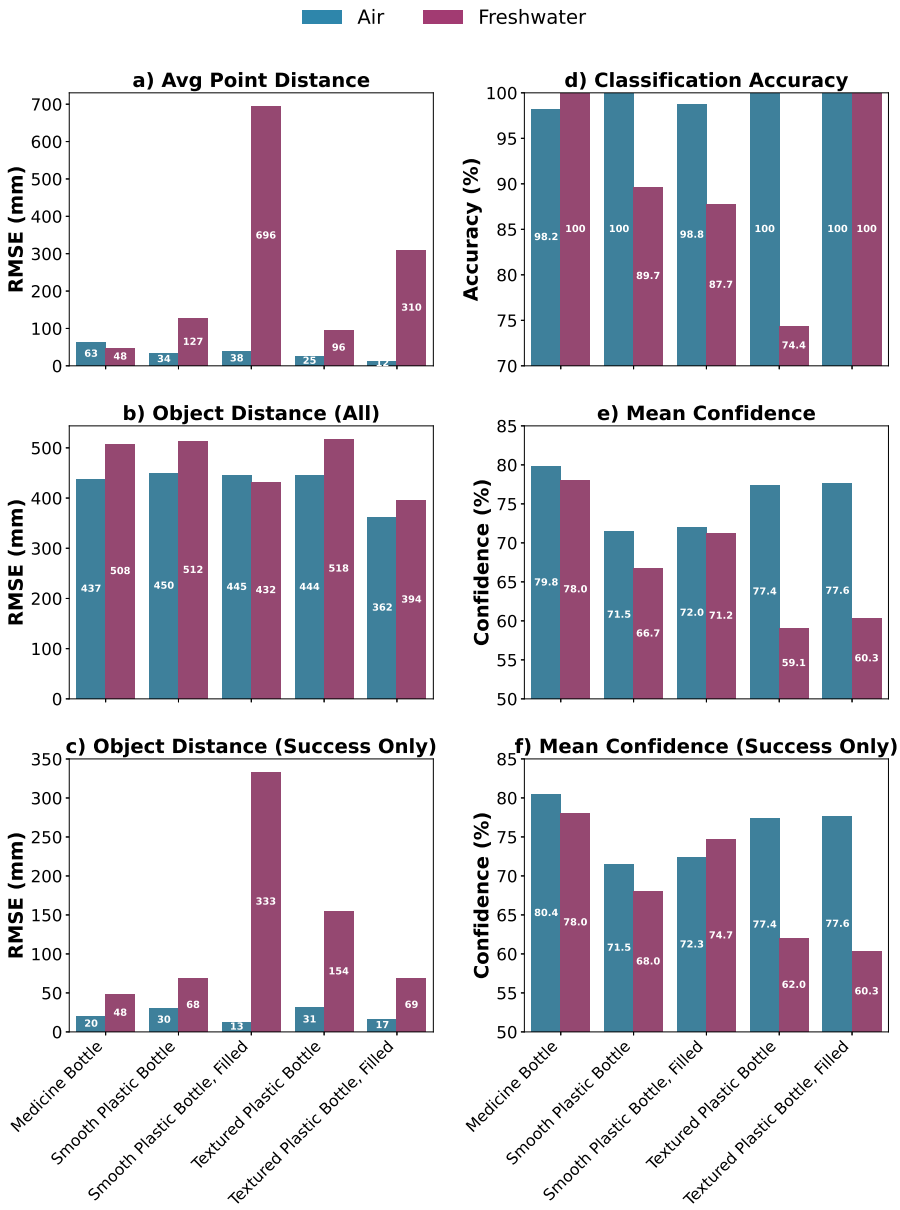

Fig. 2: Air vs Freshwater by bottle type. Panels (top→bottom): **(a)** RMSE (Average Point Distance), **(b)** RMSE (Object Distance, all), **(c)** RMSE (Object Distance, success-only), **(d)** Classification accuracy, **(e)** Mean confidence (all), **(f)** Mean confidence (success-only). Freshwater increases distance error and reduces accuracy/confidence, with the largest degradations for filled transparent bottles and the most stable behaviour for textured empty bottles.

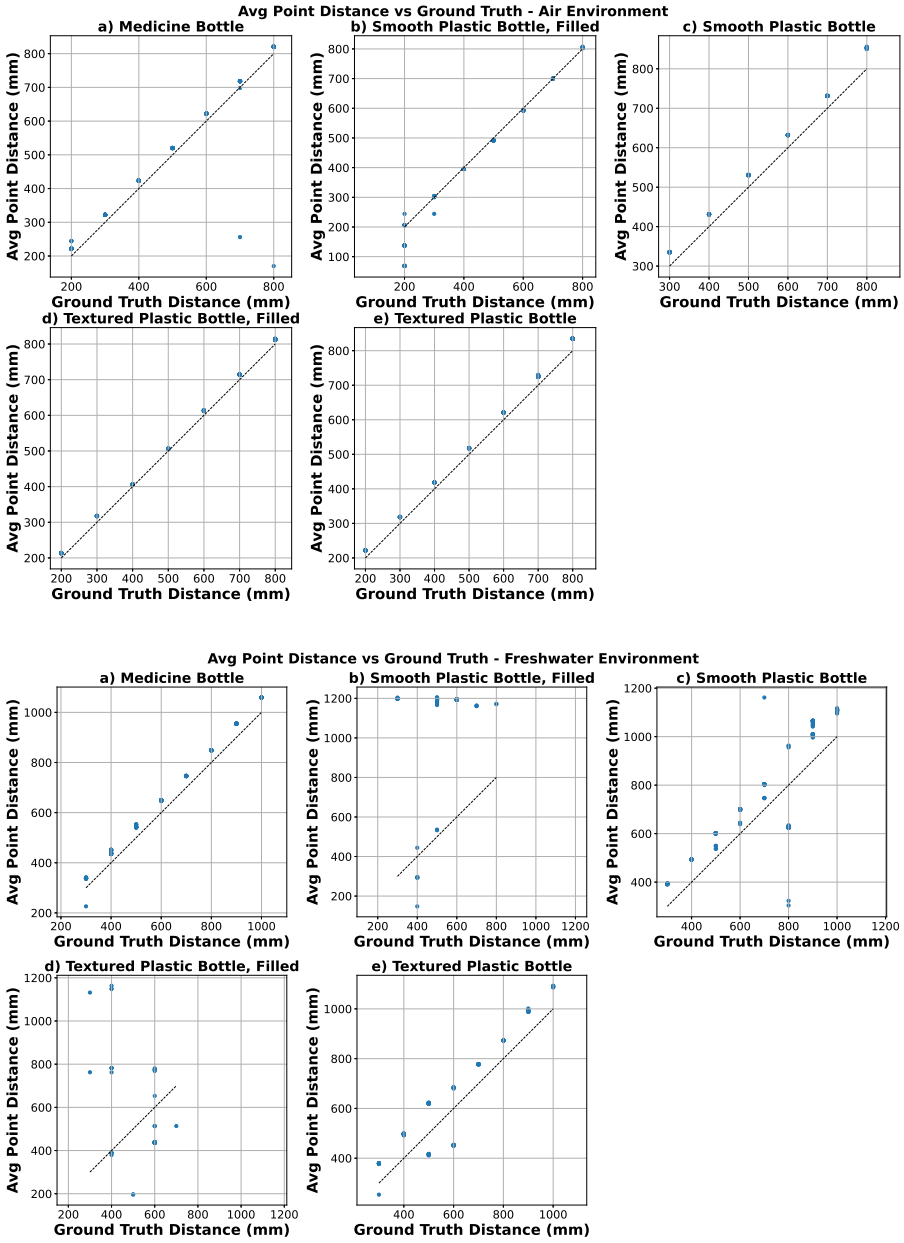

Fig. 3: Average of all three point distances across vertical samples (Upper, Centre, Lower) for Air (top) and Freshwater (bottom). Freshwater increases variability and occasional failures, especially for filled transparent bottles.

medicine (20→48 mm), smooth empty (30→68 mm), textured empty (31→154 mm), textured filled (17→69 mm), and a dramatic increase for the filled smooth bottle (13→333 mm). **Range trend.** In the per-distance plots (Fig. 4), errors grow with range and are consistently higher underwater, most notably for the filled smooth bottle, while medicine and smooth empty remain comparatively less affected. **Practical implication.** Applying a success-only or confidence-based filter improves robustness, but refractive cases (filled transparent) remain the dominant source of object-level range error underwater.

3.4 Classifier Score–Distance Relationship

Accuracy (Fig. 2d). Accuracy drops underwater for most types: smooth empty 100→89.7%, smooth filled 98.8→87.7%, textured empty 100→74.4%. Medicine and textured filled remain at 100%. **Mean confidence, all detections (Fig. 2e).** Confidence is systematically lower underwater: medicine 79.8→78.0%, smooth empty 71.5→66.7%, smooth filled 72.0→71.2%, textured empty 77.4→59.1%, textured filled 77.6→60.3%. **Mean confidence, success-only (Fig. 2f).** Filtering to correct detections raises confidence but preserves the gap for most types: medicine 80.4→78.0%, smooth empty 71.5→68.0%, textured empty 77.4→62.0%, textured filled 77.6→60.3%; smooth filled is slightly higher underwater (72.3→74.7%). **Score–distance trends (Fig. 5).** Confidence decreases with distance and is shifted downward underwater, with wider low-confidence tails—consistent with (d–f) and most pronounced for refractive (filled transparent) cases. **Gating policy.** A modest confidence threshold, optionally combined with a success-only constraint, removes low-quality detections and stabilises range estimates without changing detector weights. This mitigates—but does not eliminate—the underwater gap, especially for filled transparent bottles. **Takeaway.** Underwater conditions reduce classifier certainty and, for most types, accuracy; even after filtering, confidence remains lower for the hardest (refractive) cases.

3.5 Pre-processing Ablation: *basicEnhance*.

(a) Distance RMSE. The *basicEnhance* technique increases RMSE across vertical samples and object distance: Upper 266.9→350.9 mm, Centre 280.7→346.2 mm, Lower 274.8→340.4 mm, Object 146.2→209.3 mm (Fig. 6a). This likely reflects edge/contrast amplification that also strengthens non-target structure. **(b) Success rate.** Object-distance success improves from 41.6% to 50.8% (Fig. 6b), indicating more frequent valid range estimates. **(c) Classification accuracy.** Accuracy drops slightly, 90.3%→87.6% (Fig. 6c).

Takeaway. Technique *basicEnhance* trades a higher success rate for worse distance accuracy and slightly lower classification accuracy. It can be useful as a lightweight way to stabilise detection presence but not to improve range accuracy. Larger gains will require geometry-aware methods [1,2], but initially we suggest *confidence-gated pooling* via the ZED confidence map (CGP), *contrast-limited adaptive histogram equalisation (CLAHE) + Grey-World* (CGW), and a

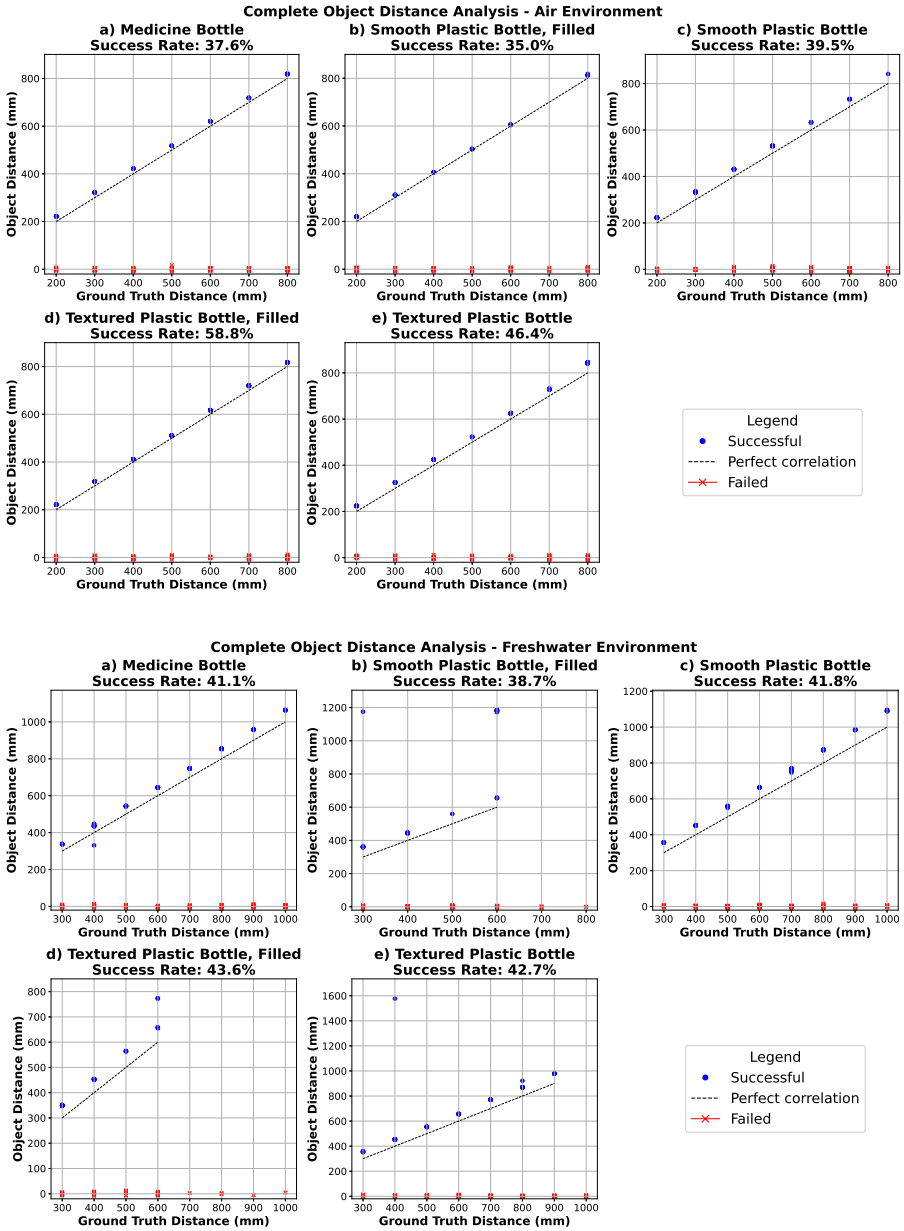

Fig. 4: Object Distance samples for Air (top) and Freshwater (bottom), including failed detections and overall success rates. Water degrades object-level distance accuracy most prominently for filled transparent bottles.

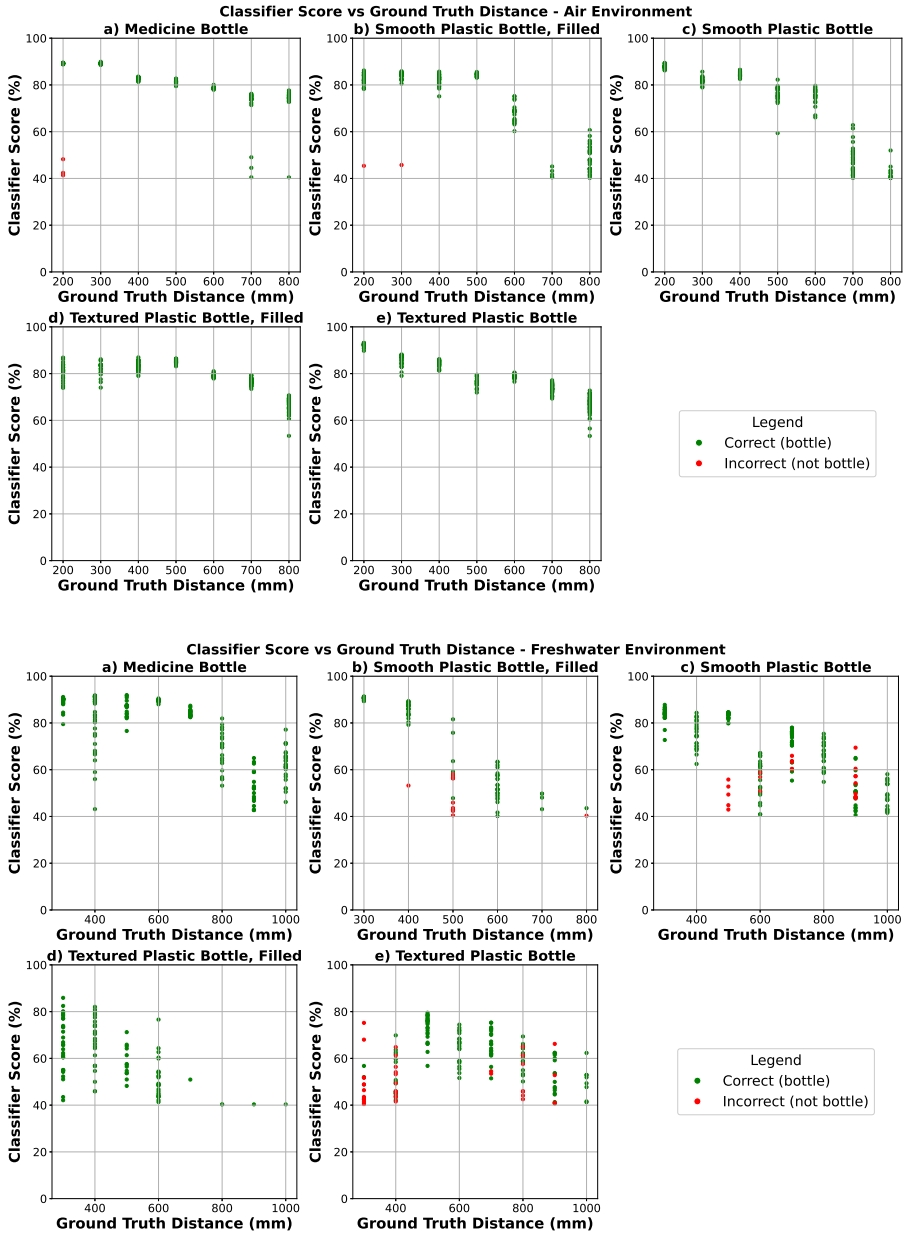

Fig. 5: Detector confidence versus ground-truth distance for Air (top) and Freshwater (bottom). Confidence levels are reduced underwater, aligning with the grouped-bar trends in accuracy and mean confidence.

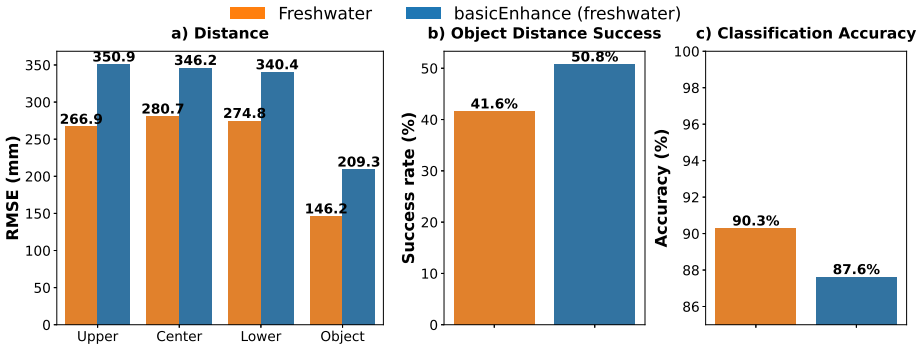

Fig. 6: Ablation on freshwater subset: baseline vs *basicEnhance* technique. Near-range detection stabilises modestly; distance RMSE changes are limited and scene-dependent.

temporal exponentially weighted moving average (EWMA) filter to reduce flicker and stabilise distance estimates [11,12,13,14].

3.6 Dataset (summary and cross-reference)

We use a multi-environment dataset pairing raw stereo recordings (file format `.svo2`) with detector outputs (file format `.xlsx`) under Air, Freshwater (plus an enhanced subset), Saltwater, and Saltwater+Pellets. The dataset *description and release* is separate to this RiTA manuscript; here we consume collated CSVs and per-recording spreadsheets for analysis. Please cite both this RiTA analysis paper and the dataset's DOI **(10.5281/zenodo.16753748)** when using the data.

4 Conclusion

Freshwater conditions consistently degrade performance relative to air: both point-wise and object-level range errors (RMSE) increase—most severely for filled transparent bottles—while classification accuracy and detector confidence decline. Textured empty bottles remain comparatively robust. Applying success-only or confidence-gated pooling improves stability but does not close the air–water gap, and simple visual enhancement raises detection success rates at the cost of higher RMSE and reduced accuracy. In practice, deployment should emphasise stability-oriented heuristics, but closing the gap will likely require methods explicitly accounting for geometry and refraction. The described computer vision pipeline can be integrated with underwater with underwater robot teleoperation interfaces to support human-operators in telemanipulation tasks [15].

Limitations. Air and freshwater runs were recorded at different times of day, so ambient lighting may differ slightly between conditions. In future work we will perform similar time analyses to quantify any time-of-day effect on detection and range estimates.

Future work. We aim to publish the dataset collected during this research. Guided by the findings in this paper, we identify follow-on implementations of further drop-in enhancements, such as: *confidence-gated pooling* using the ZED confidence map to suppress low-confidence depth (CGP), *CLAHE+Grey-World* for contrast/colour normalisation across lighting conditions (CGW), and a *temporal consistency* (TC) filter based on an exponentially weighted moving average [11,12,13,14]. While CGP and CGW target appearance and depth-quality robustness and TC adds short-horizon temporal smoothing, we expect the largest gains on filled refractive targets to come from *geometry-aware* modelling of refraction and 3D constraints (e.g., ClearGrasp and GDR-Net [1,2]), and from stronger temporal fusion beyond simple smoothing.

Acknowledgments. Aaron Smiles was funded by the UKRI EPSRC EngD Data-Centric Engineering CDT at Queen Mary University of London (reference 2601988). The work was partially co-funded by the UKRI EPSRC Q-Arena grant EP/V035304/1.

References

1. Sajjan, S.S., et al.: ClearGrasp: 3D shape estimation of transparent objects for manipulation. In: ICRA (2020).
2. Wang, G., Manhardt, F., Tombari, F., Ji, X.: GDR-Net: Geometry-guided direct regression for monocular 6D object pose. In: CVPR (2021).
3. Anonymous: Enhancing transparent object pose estimation: A fusion of GDR-Net and edge detection. arXiv preprint (2025).
4. Xie, E., et al.: Trans2Seg: Transformer for transparent object segmentation. In: IJCAI (2021).
5. Lin, Y.S., et al.: Rich context aggregation with reflection prior for glass segmentation. In: ICCV (2021).
6. Wang, Y.R., et al.: MVTrans: Multi-view perception of transparent objects. In: ICRA (2023).
7. Hsu, C.J., et al.: FakeMix augmentation improves transparent object detection. arXiv preprint (2023).
8. Smiles, A.L., Chavanakunakorn, K.D., Omarali, B., Oh, C., Farkhatdinov, I.: Implementation of a stereo vision system for a mixed reality robot teleoperation simulator. Proceedings of Towards Autonomous Robotic Systems (TAROS) conference, Lecture Notes in Computer Science, pp. 494–502. Springer, Cham (2023).
9. Smiles, A., Oh, C., Farkhatdinov, I.: Robotic perception of underwater plastic bottles for augmented telepresence. In: European Robotics Forum (ERF). Springer, Cham (2025).
10. Ultralytics: YOLOv8. Technical documentation (2023).
11. Stereolabs: ZED SDK. Technical documentation (2025).
12. Pizer, S.M., Amburn, E.P., Austin, J.D., et al.: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing **39**(3), 355–368 (1987).
13. Buchsbaum, G.: A spatial processor model for object colour perception (the Grey-World assumption). Journal of the Franklin Institute **310**(1), 1–26 (1980).
14. Hunter, J.S.: The exponentially weighted moving average. Journal of Quality Technology **18**(4), 203–210 (1986).
15. Brown J, Farkhatdinov I, Jenkin M.: ROV teleoperation in the presence of cross-currents using soft haptics. Journal of Field Robotics. 2025 Feb 23.