Exploring Stereo-Vision-based Object Detection and Ranging in Underwater Environments

by

Aaron Lee Smiles

A dissertation submitted to

The School of Electronic Engineering and Computer Science

in partial fulfilment of the requirements of the Degree of

Doctor of Engineering

in the subject of

Computer Science

Queen Mary University of London
Mile End Road
E1 4NS, London, UK

December, 2025

Declaration

I, Aaron L Smiles, confirm that the research included in this thesis is my own work, that is duly acknowledged, and my contributions are indicated. I have also acknowledged previously published materials.

I attest that reasonable care has been exercised to ensure the originality of this work, and, to the best of my knowledge, does not break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the college has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree to any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: Aaron Smiles

Date: 20/December/2025

Primary supervisor	Secondary supervisor	Author
Dr. Ildar Farkhatdinov	**Dr. Changjae Oh**	**Aaron Lee Smiles**

Exploring Stereo-Vision-based Object Detection and Ranging in Underwater Environments

Abstract

Stereo vision offers dense depth without emitting light or sound, making it attractive for teleoperation and underwater robotics. Yet transparent and refractive objects, especially plastic bottles, remain problematic due to low texture, specularities, and medium-dependent refraction. This thesis examines how far commodity stereo pipelines can be pushed in two connected contexts: providing distance cues for an augmented telepresence interface, and detecting and ranging transparent bottles underwater.

The first part implements a stereo-informed mixed reality teleoperation system combining a stereo display, a haptic device, and a Unity-based manipulator model. An initial webcam pipeline validates distance estimates in air but shows inconsistent performance on transparent targets and incompatibilities with the distance overlays component, motivating a shift to a synchronised low-cost stereo vision sensor front-end. Underwater experiments using a custom tank and recalibrated stereo rig show that opaque and textured objects remain largely tractable, whereas water-filled transparent bottles frequently yield invalid or rear-surface depths.

The second part introduces UW-TransStereo, a controlled multi-environment underwater stereo dataset focused on transparent bottle detection and ranging. It comprises 25 paired stereo vision recordings totalling over 9,000 stereo frame pairs across air and four underwater conditions (freshwater, freshwater with enhancement, saltwater, and saltwater with suspended pellets), covering three bottle targets (two transparent plastic bottles with filled and unfilled variants, and a brown glass medicine bottle), with per-frame detections, depth estimates, over 5,000 labelled measurements across environment–bottle combinations (frames, detections, ROI samples), and analysis-ready detector spreadsheets and scripts. A baseline detection-and-ranging benchmark couples detections with depth and evaluates both manual point-distance and automatic object-distance metrics.

Across environments, performance degrades systematically from air to saltwater with suspended particles, with water-filled smooth bottles driving most failures. Textured bottles are

comparatively robust, and a simple enhancement variant improves valid-range rates at the expense of higher error. The results show that commodity stereo can deliver useful telepresence cues in air and provide structured insight underwater, but refractive targets remain a fundamental challenge. The thesis motivates refraction-aware vision models, richer environmental sweeps, and tighter integration of perception with underwater manipulation systems.

Contents

List of Figures

List of Tables

Published Works

[D1] A. L. Smiles, I. Farkhatdinov, and C. Oh. UW-TransStereo: Underwater Stereo Vision Dataset for Transparent Object Detection & Ranging[1]. Zenodo, Dec. 2025. doi: 10.5281/zenodo.16753748.

[P1] A. L. Smiles, C. Oh, and I. Farkhatdinov. Multi-Environment Stereo Dataset for Transparent Object Detection & Ranging (UW-TransStereo)[2]. In: *IEEE-DATA Descriptions* **(TBC)**

[P2] A. L. Smiles, C. Oh, and I. Farkhatdinov. A Preliminary Study on Underwater Transparent Objects Detection with Stereo Vision: Air vs Freshwater[3]. In: *Proc. 13th Int. Conf. Robot Intelligence Technology and Applications (RiTA)*, 2025.

[P3] A. L. Smiles, C. Oh, and I. Farkhatdinov. Robotic Perception of Underwater Plastic Bottles for Augmented Telepresence[4]. In: M. Huber, A. Verl, and W. Kraus (eds), *European Robotics Forum 2025*. ERF 2025. Springer Proceedings in Advanced Robotics, vol. 36. Cham: Springer, 2025. Published May 22, 2025. doi: 10.1007/978-3-031-89471-8_49.

[P4] A. L. Smiles, K. D. Chavanakunakorn, B. Omarali, C. Oh, and I. Farkhatdinov. Implementation of a Stereo Vision System for a Mixed Reality Robot Teleoperation Simulator[5]. In: *Towards Autonomous Robotic Systems (TAROS)*. LNCS. Cham: Springer, 2023, pp. 494–502.

[1] https://doi.org/10.5281/zenodo.16753748

[2] Link TBC

[3] Link TBC

[4] https://doi.org/10.1007/978-3-031-89471-8_49

[5] https://doi.org/10.1007/978-3-031-38241-3_36

Acknowledgements

I would first like to give huge thanks to my supervisors, Dr. Ildar Farkhatdinov and Dr. Changjae Oh, for their continuous support along my PhD journey.

Special thanks to my CDT team, the National Oceanography Centre, and my colleagues in ARQ Robotics Lab and EECS, especially Gabriella, Pete, Pryanka, Zimpi, Yik, and Buki, for both academic and moral support.

I must also pass on my gratitude to Anthropic, OpenAI, and Cursor whose tools I used responsibly, never to take shortcuts, but to expand my work further than would have been possible without them.

Last but certainly not least, I would like to thank my friends and family, most importantly my Mam, partner Samah, cat Cloud, and dear friends Gordana and Markus, for their unwavering support. Without them, I would not have been able to achieve any of this and I am forever grateful.

This thesis is for my grandad Joe, who would have loved telling the lads at the social club that his grandson was "Dr Smiles".

List of abbreviations

3DGS	3D Gaussian Splatting
AP	Average Precision
AR	Augmented Reality
AT	Augmented Telepresence
AUV	Autonomous Underwater Vehicle
AV	Augmented Virtuality
BK7	Borosilicate Crown Glass (type)
CLAHE	Contrast Limited Adaptive Histogram Equalization
CLI	Command-Line Interface
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
DoF	Degrees of Freedom
DOI	Digital Object Identifier
EWMA	Exponentially Weighted Moving Average
FPS	Frames Per Second
GAN	Generative Adversarial Network
GT	Ground Truth
HCI	Human-Computer Interaction
HMD	Head-Mounted Display
IMU	Inertial Measurement Unit
IPD	Interpupillary Distance
LED	Light-Emitting Diode
LUT	Look-Up Table
MAE	Mean Absolute Error
mAP	mean Average Precision
MR	Mixed Reality
NaN	Not a Number
NeRF	Neural Radiance Field
PR	Precision-Recall
RGB	Red Green Blue

RMSE	Root Mean Squared Error
ROI	Region of Interest
ROV	Remotely Operated Vehicle
RSfM	Refraction-aware Structure from Motion
RWB	Rear-Wall Bias
SA	Situational Awareness
SAD	Sum of Absolute Differences
SDK	Software Development Kit
SfM	Structure from Motion
SLAM	Simultaneous Localization and Mapping
SVO	Stereo Video Object (ZED format)
SVP	Single Viewpoint
ToF	Time of Flight
UDP	User Datagram Protocol
URDF	Unified Robot Description Format
UVMS	Underwater Vehicle Manipulator Systems
VR	Virtual Reality
WRGB	White Red Green Blue
YOLO	You Only Look Once

List of symbols

I	RGB image
$I_c(\mathbf{x})$	Pixel intensity in channel c at position $\mathbf{x}$
$J_c(\mathbf{x})$	Scene radiance in channel c at position $\mathbf{x}$
B_c	Backscatter (veiling light) term in channel c
$t_c(\mathbf{x})$	Transmission coefficient in channel c at position $\mathbf{x}$
β_c	Wavelength-dependent attenuation coefficient in channel c
$d(\mathbf{x})$	Range/distance from camera to point $\mathbf{x}$
$\mathbf{r}(\mathbf{u})$	3D ray corresponding to image point $\mathbf{u}$
$\mathbf{u}$	Image point in 2D
$\mathbf{x}$	Pixel position in image coordinates
b	Stereo baseline (distance between cameras)
f	Focal length (in pixels)
K	Camera intrinsic parameters matrix
n	Refractive index
n_{air}	Refractive index of air (≈ 1.0003)

n_{glass}	Refractive index of BK7 glass (≈ 1.5168)
n_{water}	Refractive index of water (≈ 1.333)
θ	Angle (e.g., incidence angle in Snell's law)
d	Disparity (pixel shift between stereo views)
z	Depth (distance along optical axis)
$D(\mathbf{x})$	Predicted depth at pixel $\mathbf{x}$
$D^*(\mathbf{x})$	Ground-truth (reference) depth at pixel $\mathbf{x}$
D_{bg}	Background depth (e.g., tank wall distance)
M	Segmentation mask (binary or multi-class)
M_v	Valid-pixel mask (non-NaN depth estimates)
B_L	Bounding boxes in left image
B_R	Bounding boxes in right image
ε	Tolerance or threshold parameter
Bias	Systematic error (mean deviation)
MAE	Mean absolute error
RMSE	Root mean squared error
μ	Mean value
σ	Standard deviation
ρ	Correlation coefficient or density
α	Significance level (typically 0.05)
p	p-value (statistical significance)
d	Cohen's d (effect size)
conf	Detector confidence score
conf$'$	Adjusted confidence score
d_{mm}	Distance in millimeters
T	4×4 transformation matrix
$SO(3)$	3D rotation group
$\mathbb{R}^3$	3-dimensional Euclidean space over the real numbers
RWB	Rear-wall bias (percentage locking onto background)
Valid	Valid detection rate (percentage of non-NaN estimates)
J	Jaccard index (Intersection over Union)
$\lvert\cdot\rvert$	Absolute value
$\exp(\cdot)$	Exponential function
$\sqrt{\cdot}$	Square root
$\sin(\cdot)$	Sine function

Chapter 1

Introduction

1.1 Overview

Stereo vision provides dense depth estimates from a pair of cameras and underpins many remote manipulation tasks where an operator must control a robot at a distance using only camera views and limited depth cues. Passive stereo requires no active illumination and is therefore attractive for underwater robotics, where sonar and structured light face range, resolution or power constraints. However, accurate depth recovery depends on reliable correspondence between views, which degrades when targets lack texture, exhibit specular highlights, or refract light through curved transparent surfaces. Transparent and refractive objects remain particularly challenging: curved interfaces, low-contrast regions and medium-dependent refraction violate the single-medium assumptions built into many stereo pipelines, degrading both object detection and depth estimation [74]. Underwater conditions compound these difficulties through wavelength-dependent attenuation, backscatter and refraction at housing ports, reducing contrast, shifting colour and invalidating pinhole camera models [33, 74].

A key motivation for this work arose from discussions with operators and researchers at the National Oceanography Centre (NOC), who highlighted a pervasive perceptual challenge in subsea intervention: when ROV pilots operate manipulator arms using conventional 2D video displays, they routinely misjudge both the *distance* to objects and their *size*. This underwater illusion, well-known among experienced pilots, stems from the loss of binocular depth cues when a three-dimensional scene is compressed onto a flat screen, compounded by the optical properties of the underwater environment (magnification through water and viewport optics, colour-dependent attenuation, and the absence of familiar scale references). The resulting uncertainty slows manipulation tasks, increases the risk of collisions between the manipulator and delicate targets, and demands extensive pilot training to develop compensatory heuristics. These operational challenges motivated our exploration of stereo vision as a means to restore quantitative depth information and improve operator spatial awareness. A supplementary goal was to investigate whether the same stereo pipeline could provide object dimension estimates, addressing the size-misjudgment problem and supporting applications such as marine-debris characterisation where object dimensions are scientifically relevant.

Reliable, contact-free perception is central to human–robot interaction and remote manipulation. Mixed reality (MR) interfaces that overlay spatial cues onto live camera feeds can improve operator situational awareness during teleoperation, but such overlays are only as useful as the underlying depth estimates. At the same time, the marine-litter and underwater-inspection communities increasingly require automated detection and ranging of submerged objects. Marine plastic pollution represents one of the most pressing environmental challenges of our time: while surface debris is visible and accessible, studies estimate that only approximately 1% of ocean plastic floats at the surface, with the remaining 99% residing in the water column or on the seafloor [13, 88]. Early estimates suggested at least 233,400 tonnes of large plastic items and 35,540 tonnes of microplastics float in the world's oceans [16], yet this represents merely the visible fraction of a far larger subsurface accumulation. This distribution underscores the need for underwater perception systems capable of detecting and ranging submerged debris, including transparent plastic bottles that are hard to perceive with conventional sensors [74]. These application contexts motivate a rigorous study of when and why commodity stereo succeeds or fails on transparent targets, both in air and across underwater media.

Chapter 3 addresses the first strand of this problem by developing a stereo-informed MR teleoperation interface. The aim is to provide stereo-derived distance cues for MR overlays—primarily visual and audio indicators of range—to support operator awareness during remote manipulation [65, 33]. An initial unsynchronised webcam-based stereo pipeline is implemented and validated in air, revealing that opaque textured targets yield low error while transparent bottles are substantially harder to range, with occasional null readings when correspondences fail. Critical blockers in the unsynchronised setup—transport jitter, asynchronous capture and AR plug-in constraints—motivate a transition to a synchronised stereo camera with a unified software development kit (SDK). The chapter then extends the synchronised rig to an underwater testbed, constructing a bench-top tank with a Perspex viewing window, controllable lighting and recalibrated stereo, and runs materials-versus-environment and lighting studies that quantify how medium properties and target materials jointly affect ranging performance [33].

Chapter 4 builds on these findings by introducing *UW-TransStereo*, a multi-environment underwater stereo dataset and benchmark focused on transparent bottle detection and ranging [74]. The dataset pairs raw rectified stereo recordings with per-frame detector outputs and collated analysis tables across five bottle types and multiple media conditions (air, freshwater, saltwater, saltwater with suspended pellets). A reproducible capture-and-logging pipeline is released alongside the data, together with an evaluation protocol that formalises dual range estimators (manual point distance and automatic object distance) and reports metrics under both all-detections and success-only regimes. This chapter establishes the experimental design, data organisation and methodology that enable systematic cross-environment comparisons [74].

Chapter 5 presents the comprehensive experimental results obtained by applying that evaluation protocol across six test environments. Cross-environment heatmaps and per-environment scatter plots quantify how environment, bottle type and surface properties affect distance error, valid-range rate and classification performance. Statistical significance testing confirms that, for this setup, saltwater yields modestly but significantly lower error than freshwater, while detector confidence remains similar. The analysis identifies the filled smooth (transparent) bottle as the dominant failure mode and evaluates lightweight inference-time strategies (success-only filtering, confidence gating, a simple edge-and-depth-aware enhancement), concluding that such heuristics can stabilise detection presence but do not close the air–water gap for refractive cases [74].

Taken together, these chapters address two interconnected needs. First, there is a need to understand how stereo vision can provide stable, interpretable distance cues for augmented telepresence interfaces where operators must reason about distances to remote objects. Second, there is a need to quantify how far commodity stereo pipelines can be pushed for underwater transparent-object detection and ranging, and to characterise the environment- and target-dependent failure modes that emerge when refraction, turbidity and backscatter dominate. The remainder of this thesis addresses these needs through a sequence of stereo teleoperation experiments and a data-centric study of underwater transparent bottles.

1.2 Research aims and objectives

This thesis addresses the following research aims:

- To engage with industry stakeholders and operational end-users to identify real-world perceptual challenges in ROV teleoperation and establish requirements for improved spatial awareness during subsea manipulation.
- To investigate how stereo vision can provide reliable distance cues for augmented telepresence interfaces in remote manipulation scenarios.
- To evaluate the feasibility and limitations of commodity stereo vision for detecting and ranging transparent plastic bottles underwater, including the challenging case of water-filled (index-matched) objects.
- To develop and release a multi-environment underwater stereo dataset and benchmark that enables systematic evaluation of transparent-object detection and ranging across varying media conditions.

These aims are addressed through the following objectives:

- **Literature review and problem framing.**
  - Survey the state of the art in underwater image formation, refraction-aware stereo calibration, transparent-object perception, and mixed-reality interfaces for teleoperation.
  - Identify the key failure modes (rear-wall bias, invalid disparities) that arise when stereo pipelines encounter refractive and index-matched targets underwater.
- **Stereo-informed augmented telepresence pipeline.**
  - Implement a stereo vision pipeline integrated with a 3D display, haptic device, and Unity-based mixed-reality interface for teleoperation.
  - Validate distance estimation accuracy in air across diverse object types (opaque, textured, and transparent).
  - Identify and address critical blockers in unsynchronised stereo configurations, motivating a transition to synchronised stereo with vendor SDK integration.
- **Underwater stereo ranging experiments.**
  - Construct a controlled underwater testbed with stereo camera mounting, recalibration capability, and configurable lighting.

  - Conduct systematic materials-versus-environment and lighting studies to quantify how medium properties, target materials, and illumination affect stereo ranging performance.
  - Characterise failure modes specific to water-filled transparent bottles, including disparity collapse and rear-wall locking.

- **Dataset creation and multi-environment benchmark.**
  - Design and release UW-TransStereo, a multi-environment underwater stereo dataset comprising paired stereo recordings, per-frame detection logs, and collated analysis tables across air, freshwater, saltwater, and pellet-rich conditions.
  - Formalise an evaluation protocol with dual range estimators (manual point distance and automatic object distance) and metrics that expose refractive failure modes (RMSE, Bias, Valid%, Rear-Wall Bias).
  - Conduct cross-environment statistical analysis to quantify performance degradation from air to underwater and identify dominant failure modes by bottle type and environment.

1.3 Research problem

Concretely, this thesis considers commodity stereo cameras deployed in two interconnected scenarios:

- **Augmented telepresence.** A stereo rig (initially unsynchronised webcams, later a synchronised stereo camera) provides live binocular views of a scene to a 3D stereo display. An MR interface overlays a rendered manipulator model onto the stereo feed and exposes distance cues—visual and audio—derived from the stereo depth between the end effector and physical objects [65, 33]. The goal is to enable a human operator to manipulate objects at a distance with improved spatial awareness.
- **Bench-top underwater perception.** The same synchronised stereo camera is mounted behind a Perspex window on a custom experimental tank. Plastic bottles of different shapes and fill states are placed at known distances on the tank floor across air and underwater media (freshwater, saltwater, saltwater with suspended pellets), under varied lighting. A detection-and-ranging pipeline combines a modern object detector with stereo depth to estimate per-object distances and logs both successes and failures [74].

In both scenarios, the stereo system performs calibration, rectification, dense matching and triangulation to obtain per-pixel depth, with object distances derived from depth measurements within regions of interest [1, 33]. Performance is assessed via range error metrics such as mean absolute error (MAE) and RMSE, together with the valid-range rate (the fraction of frames with non-NaN depth estimates under the given protocol) [74]. The central research problem is therefore:

> *How can commodity stereo vision be used to provide reliable distance cues for augmented telepresence, and what are the behaviour and limitations of such pipelines when extended to underwater detection and ranging of transparent plastic bottles across different media and lighting conditions?*

This problem is addressed through the following research objectives:

1. **Stereo-informed augmented telepresence (Chapter 3).** Investigate how stereo-derived depth measurements can be exposed to an operator via MR overlays and sonification in a teleoperation interface, and quantify their behaviour for a range of opaque and transparent objects in air. This includes implementing a webcam-based stereo pipeline integrated with a rendering environment, validating range accuracy for diverse objects and identifying limitations that motivate a synchronised stereo front-end [33].

2. **Underwater stereo ranging for plastic bottles (Chapters 3 and 4).** Characterise the impact of medium and illumination on stereo ranging for plastic bottles by constructing an underwater testbed and running systematic materials–environment and lighting studies. This entails recalibrating the camera behind a Perspex window, deploying a controlled set of bottle types and environments, and analysing how water properties, suspended particles and coloured lighting affect error and valid-range rate [74].

3. **Dataset and benchmark for underwater transparent-object ranging (Chapters 4 and 5).** Design, release and analyse UW-TransStereo, a multi-environment underwater stereo dataset and detection-and-ranging benchmark for transparent bottles. This includes a reproducible capture-and-logging pipeline, collated tables that integrate ground-truth distances, stereo-derived distances and detector confidences, and a baseline protocol that uses both manual point-distance and automatic object-distance metrics. An illustrative edge- and depth-aware pre-processing variant (basicEnhance) is provided as a proof-of-concept

demonstrating how the dataset can be used to evaluate inference-time enhancements; this variant is designed to improve detection stability rather than ranging accuracy, so the observed trade-off between improved detection success rates and increased distance error is expected by design [74].

1.4 Contributions

Based on the above research problem and objectives, this thesis makes the following contributions:

1. **Stereo-informed MR teleoperation pipeline.** The thesis develops a stereo-informed augmented telepresence pipeline that integrates commodity stereo cameras with a 3D PluraView display, a haptic device and an MR-based robot model. The system overlays distance cues between a simulated end effector and physical targets in the stereo view, and includes an audio sonification channel whose pitch increases as the end effector approaches the object [65]. In the unsynchronised webcam configuration, the stereo system is calibrated and validated across 45–130 cm for a diverse object set, revealing that transparent objects such as clear bottles are substantially harder to range than opaque textured targets [33]. The analysis identifies critical blockers—including asynchronous capture vs rendering and AR plug-in constraints—that motivate a transition to a synchronised stereo front-end with a unified SDK for MR overlays [33]. Parts of this work have been disseminated in joint publications [P3,P4].

2. **Controlled underwater stereo testbed and analysis.** The thesis establishes a controlled underwater stereo testbed by constructing a bench-top tank with a Perspex viewing window, custom stereo camera mount and multicolour LED lighting, followed by underwater recalibration achieving sub-pixel reprojection error [33]. A materials–environment protocol is introduced in which five bottle types are placed at marked distances in air, freshwater, saltwater and saltwater with suspended pellets. Systematic analyses show that underwater conditions increase distance error and reduce valid-range rate relative to air, with suspended pellets further degrading contrast and producing more mismatches and NaNs, particularly in upper regions of the field of view [74]. Water-filled transparent bottles emerge as the hardest cases across media, often yielding NaNs or rear-surface ranges when the bottle is near index-matched to the water [74]. These studies provide a first quantitative picture of

how medium properties and target materials jointly affect stereo ranging for underwater plastic bottles. Elements of this contribution have appeared in prior work [P2,P3].

3. **UW-TransStereo dataset and benchmark for transparent bottle detection and ranging.** Building on the tank and pipeline, the thesis introduces UW-TransStereo, a multi-environment underwater stereo dataset and benchmark for transparent bottle detection and ranging [74]. The archive contains 25 paired stereo recordings (50 `.svo2` files) with accompanying per-frame detection spreadsheets and qualitative screenshots, spanning Air, Freshwater (night), Freshwater-basicEnhance, Saltwater and Saltwater-pellets [74]. Collated analysis tables integrate more than 5,000 labelled measurements across environment and bottle combinations, including ground-truth distances, stereo-derived distances at multiple regions of interest, 3D object dimensions (width, height, depth), detector confidences and condition identifiers [74]. A baseline detection-and-ranging pipeline is released alongside scripts that regenerate environment- and bottle-specific error, confidence and valid-range summaries, including Rear-Wall Bias (RWB) to quantify the index-matching failure mode whereby stereo locks onto the tank wall rather than the transparent bottle surface. The dataset also supports evaluation of dimension accuracy, revealing similar success rates to ranging but with an 11% gross measurement error rate. An illustrative basicEnhance variant that blends edge emphasis and depth-aware confidence reweighting without retraining is also provided as a proof-of-concept for evaluating inference-time enhancements against the benchmark [74]. This data-centric contribution provides a reusable benchmark and toolchain for researchers investigating underwater stereo and transparent-object perception. The dataset and associated analyses are documented in joint publications [P1,P2].

1.5 Novelty

This thesis makes the following novel contributions to the field:

- **First systematic study of commodity stereo on underwater transparent bottles.** While stereo vision has been applied to underwater perception, existing work focuses on opaque targets or does not address the specific challenge of water-filled transparent plastics. This thesis provides the first comprehensive characterisation of when and why commodity stereo fails on index-matched transparent containers underwater, quantifying the rear-wall bias phenomenon where disparity collapse causes the stereo system to triangulate the background instead of the target surface.

- **UW-TransStereo: a purpose-built multi-environment dataset and benchmark.** Existing underwater datasets lack paired stereo recordings with ground-truth depth for transparent objects. UW-TransStereo addresses this gap by providing raw rectified stereo, per-frame detection outputs, and structured analysis tables across six test conditions (Air, Freshwater Day, Freshwater Night, Freshwater with enhancement, Saltwater, Saltwater with Pellets). The dataset enables reproducible benchmarking of both classical stereo methods and learning-based approaches on refractive underwater targets.
- **Dual-estimator evaluation protocol for transparent-object ranging.** The thesis introduces an evaluation methodology that compares manual point-distance sampling with automatic SDK-based object distance, reported under both all-detections and success-only regimes. This dual-estimator, dual-regime approach exposes whether performance differences arise from estimation quality or from systematic failures (NaNs, rear-wall locking), providing actionable guidance for system deployment.
- **Quantitative evidence for saltwater advantage.** The cross-environment statistical analysis reveals a counter-intuitive finding: saltwater yields significantly lower distance error and higher success rates than freshwater for this stereo configuration, despite greater optical complexity. This result, supported by formal significance testing, challenges assumptions about underwater stereo degradation and motivates further investigation of medium-dependent stereo behaviour.
- **First characterisation of stereo-derived object dimensions underwater.** Beyond distance estimation, the thesis evaluates the ZED SDK's 3D object sizing capability (width, height, depth) for transparent bottles across environments. The analysis reveals that dimension measurements share similar success rates ($\sim$44%) with distance estimation, achieve 5–8% Mean Absolute Percentage Error (MAPE) when successful, but exhibit an 11% gross measurement error rate where dimensions deviate by more than 2× from ground truth. This complements the ranging analysis by characterising the full 3D pose estimation capability of commodity stereo for underwater transparent objects.
- **Integration of stereo ranging with mixed-reality teleoperation.** The thesis demonstrates how stereo-derived distance cues can be surfaced through visual and audio overlays in an augmented telepresence interface, linking perception outputs to operator aids. The progres-

sion from unsynchronised to synchronised stereo documents practical blockers and solutions for real-time MR integration.

1.6 Benefits to the community

This research provides the following benefits to the academic and practitioner communities:

- **Open dataset for underwater perception research.** UW-TransStereo is publicly released on Zenodo with a persistent DOI (10.5281/zenodo.16753748), providing researchers with paired stereo recordings, structured logs, and over 5,000 labelled measurements. The dataset supports algorithm development for underwater stereo, transparent-object detection, and depth completion, filling a gap in available benchmarks for refractive underwater targets.
- **Reproducible baseline and evaluation scripts.** Accompanying code enables exact regeneration of all reported metrics, heatmaps, and statistical tests. Researchers can directly compare new methods against the baseline results and extend the analysis to additional conditions or algorithms without re-implementing the evaluation pipeline.
- **Practical guidance for ROV system designers.** The systematic characterisation of failure modes—identifying water-filled smooth bottles as the dominant failure case, quantifying rear-wall bias, and documenting lighting effects—provides actionable design guidance for underwater perception systems. Engineers can use these findings to set appropriate confidence thresholds, select target materials for testing, and anticipate performance envelopes in deployment.
- **Foundation for marine-litter monitoring applications.** Transparent plastic debris, particularly bottles, is a primary target for marine pollution monitoring and cleanup operations. This work establishes baseline detection and ranging performance for such targets under realistic underwater conditions, supporting the development of automated debris collection systems and informing expectations for perception reliability in ROV-based cleanup missions.
- **Template for multi-environment dataset design.** The data organisation, naming conventions, logging pipeline, and evaluation protocol documented in this thesis provide a template for future underwater perception datasets. The approach of pairing raw sensor data

with structured per-frame logs and collated analysis tables balances archival completeness with usability for systematic analysis.

- **Bridge between telepresence and underwater perception.** By demonstrating stereo-informed MR overlays and then characterising the limits of that stereo backbone underwater, this work connects two research communities: those developing operator interfaces for remote manipulation, and those advancing underwater computer vision. The findings motivate tighter integration of perception reliability metrics with interface design decisions.

1.7 Organisation of the thesis

The remainder of this thesis is organised as follows.

Chapter 2 *Background and related work* reviews the literature on human-to-robot handovers, hand–object reconstruction, robot control for grasping and safety-aware interaction. It discusses marker-based and markerless perception approaches, the challenges of reconstructing and tracking transparent containers, and the role of simulation in de-risking handover behaviours before deployment.

Chapter 3 *Stereo Vision for Augmented Telepresence and Underwater Object Ranging* presents the implementation of the MR teleoperation pipeline and its evolution from an unsynchronised webcam-based stereo system to a synchronised stereo front-end. The chapter details the stereo calibration and depth-estimation methodology, the stereo display and haptic interface, validation experiments in air and the construction of an underwater tank and protocol for plastic bottle ranging.

Chapter 4 *UW-TransStereo: a multi-environment underwater stereo dataset for transparent bottle detection and ranging* introduces the UW-TransStereo dataset, the associated detection-and-ranging pipeline and the structure of the released archive. It describes the experimental design, tank setup, five-bottle target set, acquisition pipeline, data logging and organisation, and formalises the evaluation methodology including dual range estimators (manual point distance and automatic object distance) and metrics (RMSE, success rates, classification accuracy, detector confidence) under all-detections and success-only regimes.

Chapter 5 *UW-TransStereo: Experimental Results and Cross-Environment Analysis* presents the comprehensive results of applying the evaluation protocol across six test environments

(Air, Freshwater Day, Freshwater Night, Freshwater with basicEnhance, Saltwater, Saltwater with Pellets). It examines how environment, bottle type and surface properties affect distance error, valid-range rate and classification performance through cross-environment heatmaps, per-environment scatter plots, and statistical significance testing, identifying key failure modes and practical deployment guidance.

Chapter 6 *Conclusion* summarises the main findings and contributions of the thesis, reflects on limitations of the current stereo telepresence and underwater perception pipelines, and outlines future directions for improving geometry-aware models, integrating learning-based methods and closing the loop between augmented telepresence interfaces and underwater datasets.

Chapter 2

Background

2.1 Introduction

In this chapter, we review the state of the art that underpins vision-based underwater detection and ranging, focusing on how optical physics, refraction, and transparency jointly challenge stereo perception. The goal is to establish the technical background for our thesis problem: *stereo-based detection and ranging of transparent, bottle-like objects underwater*, where refraction and index matching frequently invalidate pinhole assumptions and cause disparity failure.

The chapter proceeds as follows. Section 2.2 defines the problem and scope—summarising the operational context of underwater teleoperation and manipulation, and why reliable near-field perception remains a bottleneck for mixed-reality (MR) operator aids. Section 2.3 introduces how MR-compatible overlays can be driven by stereo, describing the interface concepts that motivate this chapter's focus on robust depth recovery.

Sections 2.4 and 2.5 examine the *image-formation physics* governing underwater vision—attenuation, backscatter, and refraction through flat ports—and synthesise *algorithmic responses* including enhancement, learning-based stereo, and physics-informed neural rendering. These sections clarify what modern methods can and cannot fix when physics violates pinhole assumptions.

Section 2.6 formalises *refraction-aware camera models and calibration strategies*, outlining how conventional rectification must be adapted for multi-layer media and what practical recipes yield usable disparity under water. This includes calibration, rectification, and verification workflows applicable to small-ROV stereo rigs.

Section 2.7 transitions from general underwater vision to the specific challenge of *transparent and refractive-object perception*. We compare in-air transparency mechanisms (e.g., specular/refractive cue exploitation, multi-view aggregation) to their underwater counterparts and identify transferable priors for boundary emphasis and confidence gating in refractive media.

Section 2.8 surveys *datasets and evaluation protocols*, highlighting gaps in transparent-object coverage and proposing metrics that expose refractive failure modes—such as the proportion of valid disparities and rear-wall bias—needed for reproducible benchmarking.

Together, these sections provide the theoretical and empirical foundation for the subsequent methodological chapters, linking the physics of underwater imaging with calibration practice, perception reliability, and the design of operator aids that depend on those perception outputs.

2.2 Problem and scope

This section situates the thesis problem—stereo-based detection and ranging of transparent, bottle-like objects underwater—within the broader context of ROV teleoperation and manipulation. We first describe why near-field perception is critical for intervention, then narrow to the specific challenges posed by transparent targets in visually degraded conditions.

Teleoperated underwater work is commonly performed with ROVs in visually challenging conditions (attenuation, backscatter, colour shifts, and refraction through flat ports). Offshore ROV pilots balance navigation and station keeping with tool and manipulator control, with the latter consistently reported as the dominant contributor to cognitive load during intervention tasks [80]. Many underwater intervention tasks are still performed with ROVs in teleoperation, whereas AUVs are predominantly used for survey; autonomous manipulation remains challenging and places a premium on effective information transfer to a human supervisor [54]. Classic manipulation programmes such as *AMADEUS* demonstrate the breadth and precision of subsea sampling and tooling tasks required in practice [36].

Operator cognitive load is further elevated by peripheral concerns such as tether state and handling during close-quarters work near infrastructure. Operator performance hinges on situational awareness (SA); interface design should therefore explicitly aim to enhance SA [15]. Empirical studies show that immersive interfaces can improve SA for multi-robot control, supporting the case for richer perceptual, task-relevant cues (e.g., range-to-target, ghosted contours, colour-coded confidence) in teleoperation [66]. Mixed-reality visual overlays/aids that organise multi-sensor information have been reported to reduce operator load and increase SA, accelerating decision-making during remote-vehicle operation [48]—but such overlays are only as reliable as the underlying perception.

For intervention, precise end-effector control is essential and methods that reduce operator burden are beneficial [72]. Accurate near-field range perception and cues are a key enabler for grasping and alignment: binocular-vision systems for underwater targets achieve higher measurement accuracy within suitable working distances, directly supporting close-range manipulation [30]. Similarly, stereo sensing has been integrated for underwater pipe manipulation, underscoring the operational need for reliable depth cues in intervention scenarios [50, 79]. In practice, operators benefit from immediate, localised cues in the manipulator camera view (e.g., distance-to-contact indicators and aim/approach guidance) to reduce hesitation before grasp, avoid incidental collisions, and shorten the approach phase [50].

This chapter surveys the state of the art most relevant to our thesis focus: **stereo-based detection and ranging of transparent, bottle-like objects under water**, where refraction and index matching (water-filled targets) frequently break conventional pinhole assumptions and cause disparity failure modes.

Scope. We restrict attention to passive, vision-first methods that can be deployed on small ROVs without adding heavy active sensors. We emphasise: (i) underwater image formation and refraction that alter geometry and appearance; (ii) refraction-aware camera models, calibration, and their impact on stereo rectification; (iii) transparent/refractive-object perception mechanisms that *transfer* from air to water; and (iv) datasets and evaluation protocols that make failure modes and improvements measurable. Alternatives (e.g., ToF/structured light/sonar) are discussed only insofar as they justify our stereo-first approach. A full treatment of MR interface design, human factors, and telepresence taxonomies is *out of scope* here.

Thesis-specific framing. Our experimental programme quantifies when and why stereo succeeds or fails on transparent containers across environments and lighting, with special attention to the "index-matching" case (water-filled bottles) that produces NaN/rear-wall depth estimates. We use these findings to define practical calibration/rectification choices and to structure a multi-environment dataset and analysis protocol that encourages reproducible ablations. The MR/AR overlays we ultimately present (e.g., range bars and confidence shading) are thus consumers of the perception stack established in this chapter, not the subject of the review.

2.3 Mixed Reality compatible operator overlays fed by stereo

Mixed-reality (MR) visual aids that fuse multiple sensor streams into the operator's view have been shown to facilitate more intuitive robot teleoperation by structuring task-relevant information within the interface [48]. Immersive interfaces that increase embodiment and prediction can improve operator situational awareness, supporting the case for richer, spatially-registered cues over purely 2D video [66]. In underwater intervention, stereo vision has been demonstrated for real-time metric distance measurement and integrated within manipulation pipelines, making it a suitable feed for near-field MR overlays [79, 50].

2.3.1 Overlay primitives and their place in the control loop

This subsection first describes the overlay primitives that a calibrated stereo rig can generate, then situates them within common teleoperation architectures. Understanding both the *what* (the cues) and the *where* (their integration point) clarifies why stereo reliability directly gates overlay usefulness.

Overlay primitives. The following primitives can be computed from stereo and projected into the manipulator camera view to support approach and grasp:

- **Distance-to-contact readout:** disparity-derived range rendered as a numeric or bar-style cue for grasp timing [79].
- **Reticle/aim proxy:** a tool-frame aligned marker stabilised in image space; optionally sample local depth at the reticle [48].
- **Collision cones / standoff bands:** projected safety envelopes around fingertips or tools to discourage late collisions [48].
- **Confidence indicator:** a qualitative cue derived from matching costs or disparity validity to encourage hesitation when depth is unreliable [61, 62, 90].
- **Local surface patch:** a stereo-derived surface estimate to suggest approach direction [37].

Stereo-driven overlays align with known subsea manipulation needs where reliable near-field range cues are required for pipe and fixture interaction [50]. Real-time underwater stereo provides metric distance that can be surfaced directly in the view to shorten the approach phase [79]. A representative MR teleoperation view is shown in Figure 2.1, where graphical overlays

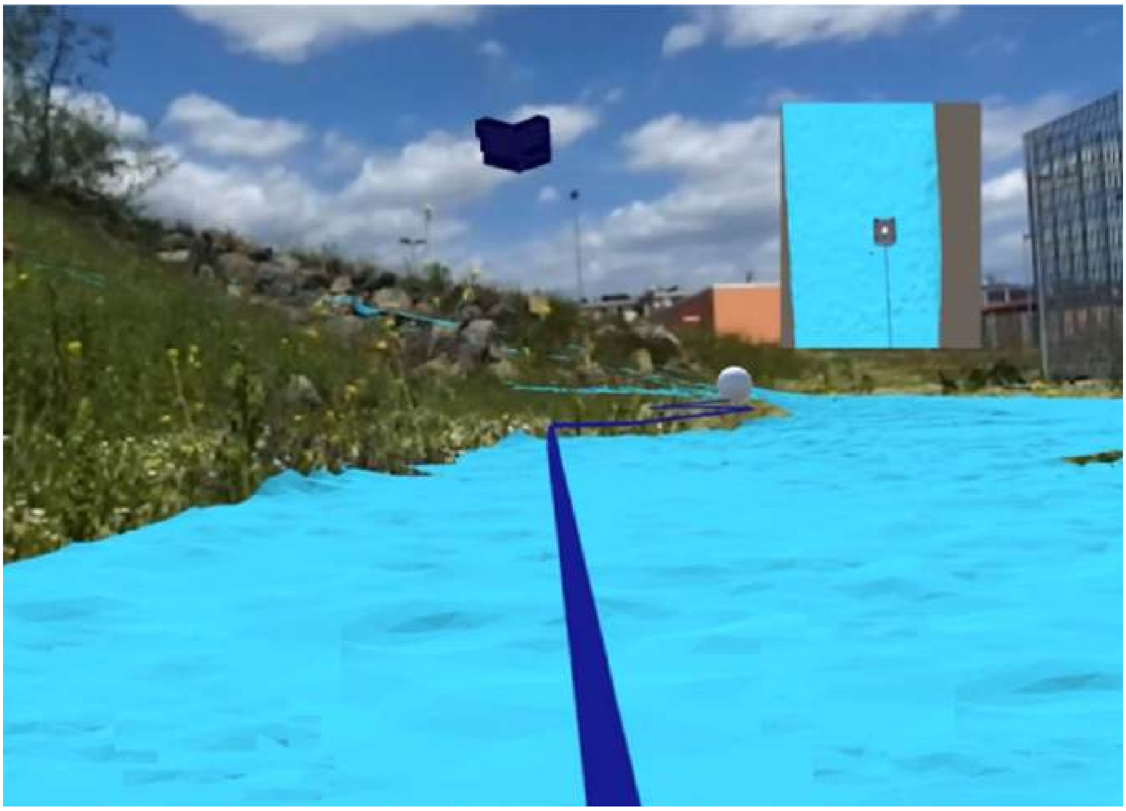


Figure 2.1: Example of an operator's MR view with live video-stream in the background (showing sky, grass, rock-stones, and buildings) and graphical representation of sensor data (suggested travel trajectory, guide arrow, top-view, and traversable area in blue). Adapted from [48], Fig. 3.

(trajectory guidance, traversable-area shading, and directional cues) are composited onto the live video stream to enhance operator situational awareness.

Situating overlays in the control loop. Where these primitives appear depends on the teleoperation mapping in use. Four common architectures illustrate the range of integration points:

Direct telepresence presents sensor imagery with minimal abstraction; overlays augment the live camera feed to enhance awareness (e.g., distance-to-contact and reticles anchored in the video) [24]. This camera-centric placement keeps cues in the operator's habitual viewpoint while adding depth and safety context [48].

Augmented telepresence (AT) denotes camera-centric telepresence enriched with spatial registration to the robot/tool frame and limited scene structure, retaining the live video as the primary canvas while stabilising overlays in 3D relative to the robot and target [6]. In AT, stereo is consumed to (i) maintain a stable reticle aligned to the tool frame, (ii) compute and display standoff distance with a confidence qualifier, and (iii) render simple collision cones that move consistently with camera/robot motion [6]. Compared with augmented virtuality (below), AT avoids a full world model and thereby reduces modelling/registration burden while still offering more predictive guidance than raw telepresence [48].

Augmented virtuality (AV) integrates live perception with symbolic affordances in a world-centric scene; such scene-centric augmentation has been shown to reduce operator collisions during remote manipulation [5]. Stereo contributes local surfaces and ranges that support guidance arrows, insertion constraints and virtual fixtures anchored in the scene [5].

Homunculus mappings place the operator inside a virtual workspace that re-embeds robot sensing and control; this approach has been demonstrated to support effective task execution in manipulation contexts [44]. Stereo supplies the near-field geometry that populates the workspace with manipulable, depth-consistent task widgets [44].

2.3.2 Implications for underwater deployment and the path forward

Underwater stereo ranging has been validated for real-time use, indicating that distance-to-contact and alignment cues can be generated within the loop of ROV manipulation [79]. Integration of stereo within autonomous manipulation on subsea vehicles further motivates exposing these depth products to human operators through MR aids to improve efficiency and safety during intervention [50, 48]. Claims regarding quantitative reductions in workload or SA gains for underwater MR overlays specifically are emerging [34]; however, immersion and predictive displays have improved situational awareness in related multi-robot teleoperation studies [66].

These operator aids—distance-to-contact, reticles and collision cones—presume a dependable near-field range backbone and a display strategy that does not overload attention [48, 66]. Feasibility evidence exists for passive stereo delivering metric distance underwater in real time and for its integration in manipulation pipelines, which makes stereo a pragmatic source for such overlays in practice [79, 50, 84]. This motivates the next body of material: core underwater computer-vision ingredients (image formation physics, calibration through refractive ports, rectification, and depth estimation) and how they affect the reliability of range and confidence cues that overlays consume.

2.4 Underwater image formation and refraction

Underwater imagery departs from in-air, pinhole assumptions due to wavelength-dependent attenuation, forward/back scattering, and refraction through flat ports. These effects alter both *appearance* (colour casts, lowered contrast, veiling light) and *geometry* (non-single-viewpoint projection), directly impacting stereo correspondence and any range estimates drawn from disparity.

2.4.1 Image formation: attenuation and backscatter

A minimal atmospheric-style *approximation* for a pixel intensity $I_c(\mathbf{x})$ in channel $c \in \{R,G,B\}$ is

$$I_c(\mathbf{x}) \;=\; J_c(\mathbf{x})\,t_c(\mathbf{x}) \;+\; B_c\big(1-t_c(\mathbf{x})\big), \qquad t_c(\mathbf{x}) = \exp\big(-\beta_c\,d(\mathbf{x})\big), \tag{2.1}$$

where J_c is the scene radiance after in-water shading, B_c the backscatter (veiling) term, β_c the wavelength-dependent attenuation coefficient, and $d(\mathbf{x})$ the range from the camera to the line-of-sight volume element. We emphasise that this is a *minimal* model; the revised underwater image-formation model shows that the direct-signal and backscatter terms follow *different* distance dependencies and coefficients, and is used by Sea-thru for accurate colour restoration [3, 4]. As depth increases, t_c decays faster for red than for green/blue, causing characteristic colour shifts; suspended particulates raise B_c, introducing a haze that reduces contrast and confounds edge-based matching.

This image-formation process is illustrated in Figure 2.2: the measured intensity I_c comprises a direct-signal component D_c (attenuated scene colour) plus a backscatter component B_c that eventually dominates at range.

Implications for correspondence. (i) Channel-dependent attenuation skews appearance priors (e.g., grayscale gradients no longer correlate across views). (ii) Range-dependent veiling raises low-frequency energy and suppresses mid–high spatial frequencies, weakening keypoint repeatability. (iii) Any pre-processing (white balance, dehazing) can introduce *nonlinear* changes that are depth-dependent and must be applied *consistently* to both views to avoid biasing stereo.

2.4.2 Refraction through flat ports: non-SVP projection

Through a flat housing port, rays traverse *air* → *glass* → *water* layers. Refraction follows Snell's law,

$$n_{\text{air}}\sin\theta_{\text{air}} \;=\; n_{\text{glass}}\sin\theta_{\text{glass}} \;=\; n_{\text{water}}\sin\theta_{\text{water}}, \tag{2.2}$$

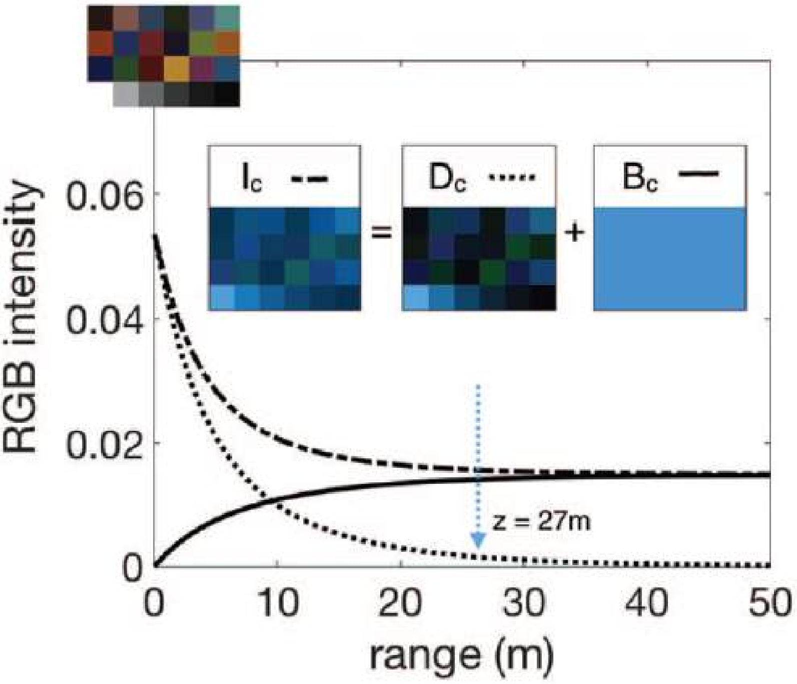


Figure 2.2: Underwater image formation is governed by an equation of the form $I_c = D_c + B_c$. D_c contains the scene with attenuated colours, and B_c is a degrading signal that strongly depends on the optical properties of the water and eventually dominates the image (shown here for a grey patch). Insets show relative magnitudes of D_c and B_c for a Macbeth chart imaged at 27 m in oceanic water. Adapted from [4], Fig. 2.

with $(n_{\text{air}}, n_{\text{glass}}, n_{\text{water}}) \approx (1.0003, 1.5168, 1.333)$ for air, BK7 glass, and liquid water at visible wavelengths (for acrylic ports, $n_{\text{acrylic}} \approx 1.49$). The resulting mapping between 3-D points and image pixels is *non-central*: effective rays do not intersect at a single viewpoint (non-SVP), invalidating the standard pinhole model [81, 2, 70].

Consequences for calibration and rectification. (i) Pinhole intrinsics estimated in air generally *do not* transfer in water behind a flat port. (ii) Epipolar lines are no longer guaranteed to be straight nor shared by a single homography; naive stereo rectification can produce biased disparities, nonuniform scale, and mismatched epipolars. (iii) Small field-of-view and near-axis rays may *approximately* admit a central model (thin-glass, near-axial approximations), but bias grows with off-axis angle, port thickness, and camera–port spacing [81, 70].

The violation of the single-viewpoint (SVP) assumption by flat-port refraction is illustrated in Figure 2.3: rays for objects at different positions appear to originate from different apparent viewpoints, making the camera system non-central (non-SVP).

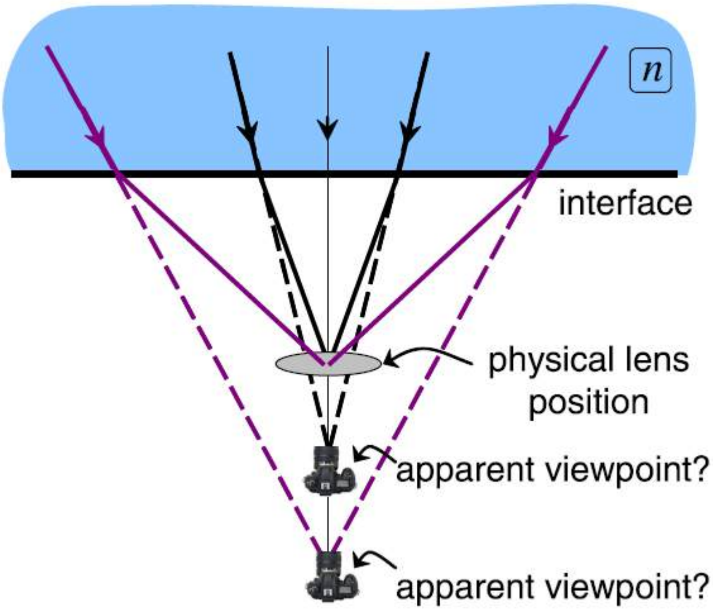


Figure 2.3: Looking through a flat interface into a medium yields a non-SVP system despite the use of a perspective camera. Rays for different scene points appear to originate from different apparent viewpoints, invalidating the pinhole model. Adapted from [81], Fig. 4.

2.4.3 Practical takeaways for stereo pipelines

This subsection summarises the practical implications of underwater image formation and flat-port refraction for a deployable stereo pipeline. The aim is to translate the physics above into actionable design choices that affect disparity validity and bias:

- **Pre-processing parity.** Apply identical, photometrically consistent corrections to both views; avoid view-specific histogram operations that distort stereo costs.
- **Geometry matters.** Prefer refraction-aware calibration (flat-port models or ray-tracing/ray-lifting) or, at minimum, *in-water* empirical re-calibration of an effective pinhole model for your working volume [70].
- **Lighting helps, selectively.** Directional, broadband (white) lighting may improve air-filled transparent targets by restoring edges/specularities, but will not resolve index-matching cases (water-filled) where disparities collapse (see §2.7).
- **Know the failure modes.** Expect *rear-wall* depth estimates and elevated %NaN disparities on water-filled transparent objects; plan evaluation to quantify these explicitly (valid-pixel ratio, bias).

Having established the physics that governs underwater image formation, the next section formalises refraction-aware camera models and their impact on stereo rectification (Section 2.6), before Section 2.7 connects these physics to transparent-object failure modes and sensing choices.

2.5 Underwater Vision: Physics-Informed Machine Learning for Depth Ranging

This section synthesises recent research on underwater depth recovery, comparing conventional stereo approaches against machine-learning and physics-guided methods. Whereas the previous section described image-formation physics, here we examine how those physics constrain algorithmic choices—and what modern methods can and cannot fix. The aim is to identify which techniques are feasible for real-time augmented telepresence on compact ROVs [32, 7, 23].

2.5.1 Why standard stereo fails: physics, geometry, and the transparent-object problem

Underwater stereo fails for two compounding reasons: appearance degradation from scattering and absorption, and geometric distortion from flat-port refraction. Light scattering, absorption, and colour aberrations present major hurdles for correspondence matching [49]. At the same time, refraction at the camera port invalidates the pinhole model; traditional perspective projection generates errors that grow with distance and off-axis angle [73]. Consequently, calibration accuracy is always expected to degrade underwater relative to equivalent air measurements [73].

The most severe operational failure arises on transparent, unstructured targets [53]. This "transparent target problem" must be solved before tackling complex challenges like real-time mapping or UVMS integration [7]. Loss of correspondence produces "NaN returns"—missing depth regions that render the target invisible in 3-D for augmented-reality overlays [82]. Even active systems such as structured light suffer this limitation: creating a point cloud requires sufficiently contrasted images [9].

Refractive calibration challenges compound these issues. If a traditional pinhole model is used, measurement accuracy degrades when the target is not close to the calibration distance [29]. Complete theoretical compensation of refraction requires an explicit physical model; implicit calibration cannot maintain high accuracy when object range varies significantly [73]. Despite these difficulties, stereo retains a theoretical advantage in achieving absolute range accuracy (millimetre-level error), provided the fundamental physical-optics issue can be resolved [22]. Monocular depth relies heavily on motion (structure-from-motion), which is problematic for stationary ROVs or fixed targets [25].

Refraction-aware geometry recovery. Achieving the highest geometric accuracy and density requires Refraction-aware Structure-from-Motion (RSfM) [58]. RSfM incorporates Snell's law and refractive camera models, performing non-linear correction to refine camera poses and 3-D point triangulation [58]. Deep learning enhances feature handling in these pipelines: networks such as SuperPoint and SuperGlue are crucial for robust correspondence matching in visually degraded or low-texture scenes [58]. Traditional detectors (e.g., SIFT) often fail on motion blur, low-contrast features, or specular surfaces [58]. However, deep-learning localisation methods lack native refraction correction and must be coupled with explicit refraction correction to maintain metric precision [58].

2.5.2 Algorithmic responses: enhancement, learning, and neural rendering

Given the physics-induced failures above, three families of algorithmic responses have emerged: image enhancement, learning-based depth, and physics-informed neural rendering. Each addresses different parts of the problem.

Image pre-processing and enhancement. Enhancement addresses low contrast, blur, and colour cast—degradations that cause stereo mismatch [89]. Traditional methods employ parameter-free enhancement using homomorphic filtering and wavelet decomposition, correcting contrast disparities without needing prior knowledge of depth or water quality [89]. Deep learning offers efficient restoration via lightweight CNN architectures (e.g., UWCNN), enabling fast processing suitable for embedded systems [39]. GANs such as WaterGAN synthesise images from in-air RGB-D pairings, training networks for real-time colour correction that models range-dependent attenuation [40].

However, conventional enhancement focusing on appearance is insufficient to solve the geometric failure for transparent objects [14]. A dedicated physical, geometry-aware model is required to fully resolve index matching [14]. Sea-thru removes water effects by leveraging a revised image-formation model, using known range (from stereo/SfM) to estimate backscatter and true colour [4]. SyreaNet integrates synthetic and real data under this revised model, employing domain adaptation for robust enhancement [87].

Learning-based detection and ranging for debris. Detecting and ranging (D&R) marine plastic debris motivates deep-learning solutions for automatic, rapid, large-scale monitoring [63]. Detection relies on efficient architectures; Tiny YOLO provides real-time performance on embedded hardware (Jetson TX2) [17]. Optimised architectures reduce model size (e.g., 5.86 M parameters) and cut FLOPs by approximately 60% [26]. For ranging, the shift is toward physics-informed neural methods; research confirms that backscatter and direct-signal coefficients differ by sensor and water type, which is critical for accurate depth recovery [3].

Neural rendering and future directions. SeaThru-NeRF extends Neural Radiance Fields using physical constraints to infer depth in scattering media [38]. SeaSplat combines 3-D Gaussian Splatting with a physically grounded image-formation model for real-time, colour-corrected geometry rendering [94]. It is essential to distinguish their purpose: Gaussian Splatting excels at photorealistic visualisation but is unsuitable for applications requiring millimetre-level metric accuracy [58]. Alternative modalities—light-field cameras coupled with Deep Convolutional Neural Fields, or opti-acoustic fusion—offer complementary pathways for geometry acquisition [51, 21].

2.5.3 Datasets, system integration, and current limits

This subsection surveys data availability and integration constraints that affect reproducibility and real-time deployment.

Dataset scarcity and domain shift. Underwater environments impose challenges due to pervasive low-texture targets, hindering reliable stereo correspondence [20]. Deep-learning models must generalise across diverse conditions; synthetic models may struggle to distinguish close colour styles (e.g., blue vs. green water) [41]. The Underwater Dataset Fitting Model (UDFM) merges diverse degraded datasets to improve generalisation [14]. A critical scarcity persists regarding large, curated underwater datasets containing ground-truth depth and comprehensive object labelling [21]. Synthetic data generation is therefore crucial: OpenWaters generates photorealistic scenes with ground-truth labels via hardware ray tracing [57]; UWNR learns degradation models from authentic images, avoiding biases from hand-crafted priors [96]. Reproducible evaluation requires standardised data such as the CADDY dataset, which provides rectified stereo images and IMU data for 3-D pipeline testing in low-texture contexts [20]. Hardware studies confirm baselines: submerged low-cost systems (ZED/Jetson TX2) achieve accuracy better than 1 cm for distances exceeding 1 m [68, 85, 82].

Integration with augmented telepresence. Stereo vision is crucial for complex robotic tasks such as grasping and manipulation in the offshore industry [49, 85]. Stereo supports hybrid

SLAM on Underwater Vehicle Manipulator Systems (UVMS), enabling real-time mapping that underpins augmented telepresence [7]. MR interfaces must enhance presence through immersion and interactivity, and successful telepresence requires careful integration of high-fidelity 3-D reconstruction [23, 89]. Standardised HCI scales (NASA-TLX, SUS, IPQ) are used to measure usability and immersion [23]. However, current systems face integration limits: high-quality visual output often requires computationally expensive optimisation (e.g., graph cuts), restricting some processes to offline application [8]. Reliance on *a priori* knowledge of cooperative targets (e.g., pipes) limits utility for unstructured debris [49]. Practical constraints also include HMD cost and user discomfort [23]. Embedded systems achieve high throughput (e.g., 10.20 MPixels/s) for real-time feedback, but often rely on simplified restoration models that ignore complex backscatter, reducing visual fidelity for AR overlays [78].

Section summary. Stereo vision faces fundamental limitations rooted in physical refraction and catastrophic ranging failure for transparent objects [53, 73]. Resolving these demands Refraction-aware SfM and physics-guided neural architectures (SeaSplat, SeaThru-NeRF) to recover metric 3-D geometry [58, 94, 38]. The critical gap remains reliable, real-time absolute ranging for unstructured plastic debris within embedded-system constraints required for advanced augmented teleoperation [7]. Future work must integrate geometric solutions with multimodal sensing and robust HCI evaluation [21, 23]. Building on this foundation, the next section formalises the refraction-aware camera models and calibration workflows that make stereo usable under water.

2.6 Refraction-aware models, calibration, and underwater stereo

Conventional stereo assumes a central (single-viewpoint) pinhole model with straight, shared epipolar lines. Behind a *flat* housing port, rays traverse air→glass→water layers and the projection becomes non-central; epipolar loci are generally *curves*, and applying in-air intrinsics leads to biased disparities. This section summarises the minimal geometry and the practical calibration/rectification choices that make stereo usable under water.

2.6.1 Camera models for flat-port imaging

Let $\mathbf{u} \in \mathbb{R}^2$ be an image point, and let $\mathbf{r}(\mathbf{u})$ denote the corresponding 3-D ray obtained by tracing through a plane-parallel, multi-layer interface using Snell's law:

$$n_{\text{air}} \sin\theta_{\text{air}} = n_{\text{glass}} \sin\theta_{\text{glass}} = n_{\text{water}} \sin\theta_{\text{water}}, \quad (n_{\text{air}}, n_{\text{glass}}, n_{\text{water}}) \approx (1.0003, 1.5168, 1.333) \text{ (air, BK7 glass, water; acrylic ports are } \approx 1.49\text{).} \tag{2.3}$$

Key parameters are the *in-air* intrinsics, the port plane equation (normal and distance from the camera), glass thickness, and the water refractive index. The resulting camera is *non-SVP*; central pinhole approximations are valid only for narrow fields-of-view and near-axis rays, with error growing off-axis and with port thickness [81, 2, 70].

Dome ports. A hemispherical port *whose center of curvature coincides with the lens entrance pupil* approximately restores central projection (SVP); misalignment degrades this benefit. This is often impractical on compact ROV rigs but is useful context when comparing housings [81, 70, 35].

2.6.2 Calibration strategies that work in practice

We distinguish three deployable options, chosen by hardware constraints and the required accuracy window:

1. **Full refractive calibration (preferred when feasible).** Estimate the port plane, camera–port spacing, and in-water intrinsics by observing a submerged calibration target (checkerboard/AprilTag grid) at multiple tilts and depths. Optimise a refractive projection that ray-traces through the layers [70, 2]. Pros: physically grounded, transferable across working distances. Cons: more setup; requires good target visibility.
2. **Effective pinhole re-calibration (working-volume approximation).** Empirically fit a standard pinhole model *in water* for the expected working distance (e.g., 0.5–2.0 m). This reduces bias *within* the fit volume but drifts outside it. Useful when you cannot fit a refractive model or when port parameters are unknown. *Implementation note:* In this thesis we used the default OpenCV calibration (central pinhole with Brown–Conrady radial+tangential distortion) on the ZED stereo pair, producing an *effective* in-water model calibrated over our intended working range; this should be interpreted as an empirical fit rather than a refractive physical model.

3. **Precomputed warps / LUTs ("virtual pinhole").** Learn a per-pixel warp from the observed refractive image to a *virtual* central model for a chosen depth plane (or a small set of planes); then run conventional calibration on the warped images. This is fast at runtime but depth-dependent; use with caution if the scene spans a wide depth range.

In all cases, calibrate both cameras consistently and record: (i) the water salinity/temperature (for n_{water}), (ii) port/housing details, and (iii) lighting used. These metadata are necessary to reproduce rectification and to interpret residual bias.

2.6.3 Stereo rectification and triangulation under refraction

Classical rectification assumes a homography that maps epipolar lines to horizontals in both images. With a flat port, epipolar *curves* depend on viewing direction. Three practical routes exist:

Ray-space rectification. For each pixel, compute its refracted 3-D ray $\mathbf{r}(\mathbf{u})$ and project to a *virtual* image plane such that corresponding rays from the left/right cameras lie on the same scanline. This yields near-horizontal epipolars and allows standard stereo matching on the rectified pair [81].

Ray-to-ray triangulation (no rectification). Match features or patches in the native images; then intersect the two refracted rays numerically (minimum-distance point). This avoids rectification but shifts complexity to matching and intersection; it is robust when only sparse correspondences are needed.

Empirical rectification (LUT-based). Learn per-pixel rectifying warps from a submerged planar target observed across the field. This is the simplest to deploy, but the rectification is depth-biased; verify residual vertical disparity across depths in your working range.

Photometric consistency. Whatever the geometric path, apply identical pre-processing to both views. View-specific white balance, tone-mapping, or dehazing will bias matching costs and increase left–right inconsistency. *Practical note:* OpenCV's standard (and fisheye) calibration models are *central* camera models and do not explicitly model flat-port refraction. When used underwater they therefore act as effective pinhole fits valid over a limited working volume; we therefore validate residual vertical disparity on submerged planar targets and report depth bias within the calibrated range.

2.6.4 Learning-assisted stereo and monocular depth

CNN stereo (cost volumes/aggregation) and monocular depth networks can improve *appearance* robustness, but they do not fix refractive *geometry*. Use them as components *after* refraction-aware rectification, or within the effective pinhole range if you validated bias. Always report left–right consistency, %valid disparities, and bias on planar targets at known distances.

2.6.5 Failure modes, diagnostics, and a practical recipe

In underwater transparent-object ranging, failure is often systematic rather than random noise. The following failure modes should be expected, measured explicitly, and used to drive confidence gating for downstream MR overlays:

- **Rear-wall bias.** On water-filled transparent objects, disparities collapse and stereo returns background ranges; report bias vs. ground truth and highlight these regions.
- **%NaN / invalid disparity.** Low-contrast, veiled regions and refractive boundaries yield NaNs; track the valid-pixel ratio as a primary metric.
- **Curved epipolars / vertical disparity residuals.** Diagnose with a submerged planar target swept across depth; plot residual vertical disparity after rectification.
- **Left–right inconsistency.** Large LR disagreement is a strong indicator of refractive mismatch; gate range overlays in MR by a confidence derived from LR checks.

Practical recipe (used in this thesis). To translate the above into a deployable workflow, we followed these steps:

1. Place a submerged calibration target at several tilts and depths across the intended working range; record salinity/temperature.
2. Fit a refractive model (or an effective pinhole if necessary) for both cameras; export intrinsics, extrinsics, and a rectification transform (ray-space or LUT).
3. Verify rectification with a planar target: report vertical disparity statistics across depth and field-of-view.
4. Run stereo with photometrically identical views; enforce LR consistency and invalidate low-confidence pixels.
5. Validate depth with known targets (opaque and transparent, both air-filled and water-filled) to characterise rear-wall bias and %NaN.

These choices make the downstream operator aids (range bars and confidence overlays) reliable consumers of disparity, while keeping the system deployable on small ROVs.

2.6.6 From stereo feasibility to transparent-object perception

While stereo supports many intervention tasks, transparent consumer plastics remain difficult for both recognition and depth recovery: commodity depth systems commonly fail on transparent surfaces, and underwater debris datasets highlight plastic bottles as a primary (yet challenging) target class [67, 92, 99, 19], yet there is a lack of literature on water-filled plastics even though they are central to pollution-removal use-cases. At the same time, the marine-litter community has begun to apply modern detectors and segmentors to underwater imagery and bottle-centric targets, indicating both operational need and available baselines [18, 64, 55]. The remainder of the review therefore narrows from general underwater perception to transparent-object detection under aquatic imaging constraints, motivating paired stereo–detection data, reproducible testing protocols, and lightweight ablations to characterise failure modes and performance envelopes.

2.7 Transparent and refractive objects: from air mechanisms to underwater

Transparent objects pose two coupled challenges: *appearance* is dominated by background remapping and specular highlights rather than surface albedo, and *geometry* is refractive, so ray paths bend at the interface. Underwater, these effects interact with attenuation and backscatter, further degrading the cues relied upon by standard detectors and stereo systems.

2.7.1 Why transparency is hard (and harder under water)

Transparent objects defeat standard stereo and detection pipelines for reasons that are physics-grounded and therefore predictable. Understanding these mechanisms guides both failure diagnosis and the design of mitigation strategies:

- **Background remapping.** Pixels on a refractive object depict the *background* refracted through the interface. Inliers for classical matching often link to the background rather than the object surface, corrupting disparity near boundaries.
- **Weak intrinsic contrast.** With little diffuse component, edges derive from refractive-flow discontinuities and specular highlights; both are view- and lighting-dependent.
- **Index matching.** Water-filled containers reduce the refractive contrast (glass $n \approx 1.52$,

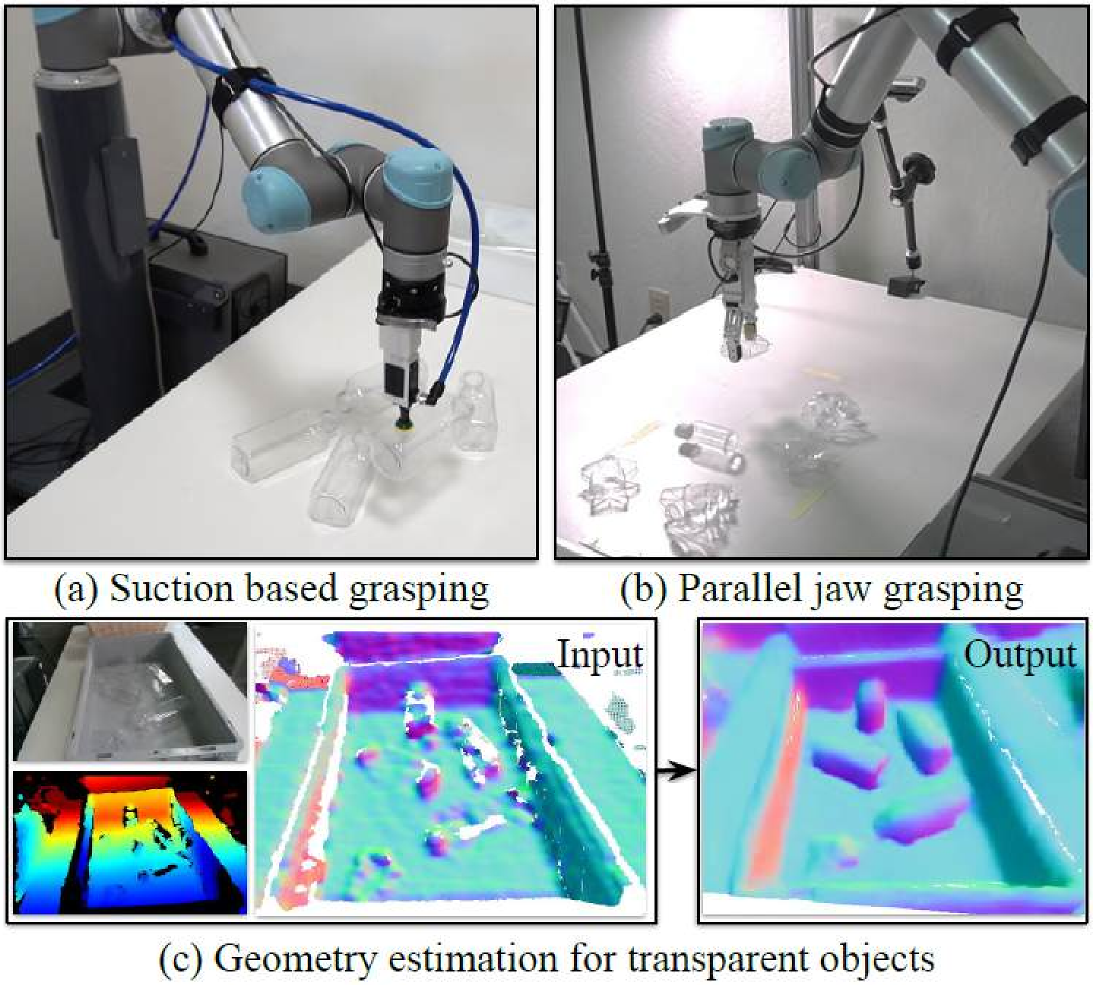

(a) Suction based grasping
(b) Parallel jaw grasping
Input
Output
(c) Geometry estimation for transparent objects

Figure 2.4: ClearGrasp leverages deep learning with synthetic training data to infer accurate 3-D geometry of transparent objects from a single RGB-D image. The estimated geometry can be directly used for downstream robotic manipulation tasks (e.g., suction and parallel-jaw grasping). Adapted from [67], Fig. 1.

water $n \approx 1.33$) [10], collapsing disparities and producing rear-wall estimates and NaNs where costs are ambiguous [31, 76].

- **Underwater formation.** Wavelength-dependent attenuation and veiling light reduce mid–high spatial frequencies, weaken boundary cues, and lower keypoint repeatability (cf. Eq. 2.1).

These challenges are illustrated in Figure 2.4: commodity depth sensors produce noisy or missing depth on transparent surfaces, motivating learning-based approaches that infer geometry from RGB cues.

2.7.2 Lessons from in-air transparency and what transfers underwater

Although underwater imaging differs substantially from in-air conditions, several mechanisms developed for transparent-object perception in air provide useful priors that can be adapted. The following in-air approaches have shown promise:

- **Segmentation via specular/refractive cues.** Networks that exploit specular highlights, boundary normals, or context (e.g., *ClearGrasp*, *Trans10K*-family) [67, 91, 92].
- **Pose/shape from multi-view.** Keypoint- or contour-based multi-view (*KeyPose*, *ClearPose*) that stabilise estimates by aggregating across views [47, 11].
- **Ray-aware rendering.** Radiance-field approaches modelling reflection/refraction (*Ref-NeRF* and variants) capture refractive flow explicitly; useful as a conceptual anchor even if overkill for real-time ROVs [83, 12].

These methods contribute priors (boundary emphasis, confidence gating, multi-view aggregation) that remain valuable underwater once photometric consistency and refraction-aware calibration are enforced (Sections 2.4–2.6).

Underwater-specific factors. Beyond adapting in-air mechanisms, several underwater-specific interventions can improve transparent-object perception:

- **Lighting.** Broad, directional white lighting can recover usable edges/specularities for *air-filled* transparents, aiding detection and stereo; it typically fails to resolve water-filled, index-matched cases.
- **Background management.** Contrasting, textured backdrops behind the workspace increase boundary salience; this helps detectors and reduces rear-wall bias during validation.
- **Polarisation.** Polariser pairs can attenuate specular glare from flat regions and offer weak surface-normal cues. Benefits drop with turbidity and multiple scattering; hardware burden is low, hence worth consideration.
- **Multi-view consistency.** Short-baseline stereo or small-viewpoint sweeps remain the most deployable route to stable pose/range on translucent objects when index matching is not severe.

Table 2.1: Modalities for transparent-object perception with representative methods and the primary signals or outcomes they leverage.

Modality	**Representative methods**	**Key signals and outcomes**
RGB only	Trans2Seg, Trans4Trans	Global context improves weak boundaries; efficient Transformer variants reach strong mIoU for safety-critical navigation [92, 97].
RGB + Reflection prior	RCAM + RRM	Multi-scale boundaries and specular cues markedly improve glass detection [42].
RGB + Polarisation	PGSNet	Fusing AoLP/DoLP with RGB yields consistent gains over strong baselines and improves robustness to illumination changes [56].
RGB + Thermal	RGB–T segmentation	Thermal–RGB fusion exposes boundaries invisible to RGB alone and outperforms single-modality models [28].
RGB + Event (underwater)	TransCODNet	Pixel-level RGB–event fusion sharpens contours and improves detection of transparently camouflaged organisms [52].
Stereo RGB	KeyPose; StereOBJ-1M baselines	Stereo keypoints achieve millimetre-level 3D accuracy; stereo improves ADD(-S) AUC by $\geq$25% over monocular on transparent/reflective objects [47, 46].
Light-field	LIT	Angular cues from light-fields enable refractive localisation and pose on ProLIT [98].
Multi-view depth	DepthSplat	Injecting monocular priors into cost volumes mitigates failures on reflective and low-texture regions [93].

2.7.3 Sensing modalities for transparency under water (synthesis)

We summarise deployable sensing choices and their underwater relevance in Table 2.2. The aim is to keep only modalities that are realistic on compact ROVs and that can feed a stereo-first pipeline.

Takeaways. RGB stereo remains the most deployable backbone; polarisation is a low-cost assist; learning-based mono detection fills gaps but cannot replace geometry; active range is constrained by turbidity and range. For water-filled, index-matched targets, failure is common—hence the need to detect and *flag* low-confidence disparity regions so operator aids consume range conservatively.

2.8 Datasets, benchmarks, and evaluation

A useful benchmark for this thesis must expose *when* stereo fails on transparent, bottle-like targets under water and quantify the gains of calibration/rectification and sensing choices. We first position existing datasets, then define an evaluation protocol tailored to refractive failure modes (rear-wall bias, %NaN).

2.8.1 Existing datasets and what they cover

Underwater object detection/segmentation. The *URPC* lineage provides annotated marine-life and object data with challenging visibility; it is widely used for underwater detection but rarely includes transparent consumer plastics or paired stereo [45]. *SeaClear* offers instance-level annotations of marine litter collected by ROVs/AUVs; it targets plastics but does not provide ground-truth depth for transparent objects [99]. *Trash-ICRA19/UMD* focuses on marine debris detection in real underwater scenes; again, transparent targets are underrepresented and depth supervision is limited [19].

Transparent-object datasets (mostly in air). *ClearGrasp*, *Trans10K*, *KeyPose*, and *ClearPose* offer segmentation and pose/depth supervision for transparent objects, but their capture conditions and image formation differ from underwater scenes; water-filled cases and refractive projection through flat ports are not addressed [67, 92, 47, 11].

Stereo/depth under water. There are few public sets with *paired stereo and depth* that also include transparent targets. Synthetic renderings exist, but they sidestep backscatter and true refraction. As a result, cross-domain transfer to field ROV footage remains fragile.

Gap. Public data under-represents water-filled, transparent containers with paired stereo and reproducible evaluation across environments (pool/tank/open-water) and lighting. This gap motivates the multi-environment, paired-stereo dataset and protocol used later in the thesis.

Representative annotated imagery from marine-litter detection work is shown in Figure 2.5, illustrating the visual challenges (illumination variation, clutter, small targets) and annotation styles typical of current underwater debris datasets.

2.8.2 Evaluation protocol for underwater transparency

We evaluate two linked tasks: (A) *detection* of transparent containers; (B) *stereo depth* within detected regions. The protocol quantifies conventional accuracy alongside refractive failure modes

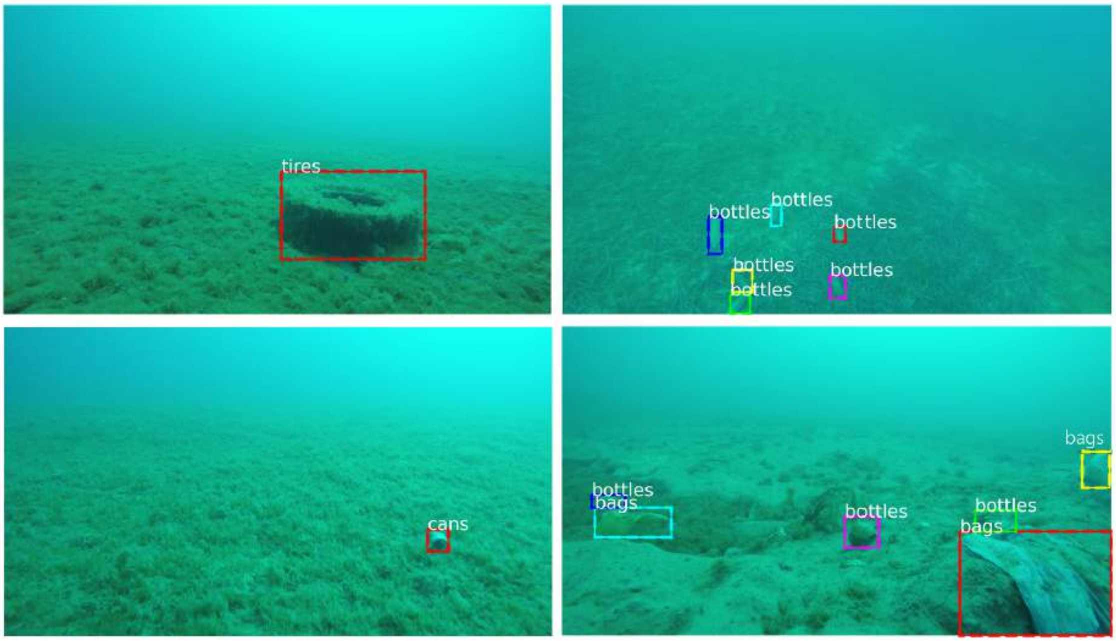


Figure 2.5: Sample images with litter objects (tires, bottles, cans, bags) and their bounding-box annotations from towed-camera marine-litter detection. Adapted from [63], Fig. 3.

specific to underwater transparent targets.

A. Detection metrics. Detection evaluation quantifies how well the system localises and classifies transparent containers. For multi-class detection benchmarks, mean Average Precision (mAP@ [.5:.95]) is the standard COCO metric, integrating precision–recall across intersection-over-union thresholds [43]. However, for single-class evaluation (e.g., bottle-only) with a fixed detector, simpler proxies suffice:

- **Classification accuracy:** proportion of detections correctly labelled as the target class.
- **Detector confidence:** mean predicted class probability, indicating model certainty.
- **Per-environment reporting:** scores stratified by environment (e.g., air, freshwater, saltwater) and lighting condition to expose domain-dependent degradation.

B. Stereo depth on transparent targets. Depth evaluation focuses on whether the stereo system returns valid, accurate range estimates within detected regions. Let D be predicted depth and D^* a reference (plane-at-known-distance or calibrated tank-wall distance). For each detection, sample depth at representative locations and compute:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(D_i - D_i^*\right)^2}, \tag{2.4}$$

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}\left(D_i - D_i^*\right), \tag{2.5}$$

where N is the number of valid depth samples. Beyond conventional accuracy, two refractive failure indicators are critical:

- **Valid% (success rate):** proportion of detections returning a non-NaN, non-zero depth estimate. High invalid rates indicate stereo matching failure, common when index matching causes the target to appear optically transparent.
- **Rear-wall bias (RWB, this thesis):** when a background plane at known distance D_{bg} exists (e.g., tank wall at 1200 mm), compute the percentage of valid depth estimates that lock onto the background rather than the target:

  $$\mathrm{RWB} = 100 \cdot \frac{|\{i : D_i > D_{\mathrm{bg}} - \varepsilon\}|}{N},$$

  with tolerance ε (e.g., 50 mm). Large RWB indicates disparity collapse—the stereo matcher sees through the water-filled transparent object and triangulates the rear wall instead. This metric directly quantifies the index-matching failure mode.

Distributional considerations. Depth estimation errors from stereo systems frequently exhibit non-Gaussian distributions due to outliers from matching failures, occlusion boundaries, and refractive artefacts [27]. Heavy-tailed and skewed error distributions are characteristic of depth sensing, particularly when transparent or low-contrast targets induce rear-wall locking or correspondence ambiguity. This motivates the use of non-parametric statistical tests (e.g., Mann-Whitney U) when comparing error distributions across conditions, as parametric alternatives assume normally distributed data and may yield unreliable inferences when this assumption is violated.

Photometric and geometric controls. Apply identical pre-processing to both stereo views; log calibration parameters (including water temperature/salinity for n_{water}), housing/port geometry, and lighting setup. Report the rectification method used (ray-space, LUT-based, or effective pinhole) to enable reproduction.

2.8.3 Reporting template (minimal, reproducible)

To make results comparable across environments and calibration choices, we report a minimal set of metrics and metadata in a standardised template. This ensures the key refractive failure modes (missing depth and rear-wall locking) remain visible in summary tables:

- **Per-environment table:** *Detection:* classification accuracy, mean detector confidence; *Depth:* RMSE, Bias, Valid%, RWB%. Stratify by bottle type (e.g., opaque, translucent, water-filled) and environment to expose target- and condition-specific failure modes.
- **Ablations:** where resources permit, vary one factor at a time: (i) calibration choice (refractive vs effective pinhole), (ii) rectification strategy, (iii) lighting (ambient vs directional), (iv) pre-processing (e.g., edge-enhanced). Report corresponding changes in Valid% and RWB to isolate effects.
- **Confidence gating:** report detection yield at representative confidence thresholds (e.g., 50%, 60%) to inform operator-aid policies that suppress low-confidence overlays.

Quality metrics for enhancement (optional). If an enhancement stage is applied before stereo, objective metrics such as UIQM or UCIQE can supplement qualitative claims [59, 95]; ensure enhancement is applied *identically* to both stereo views to avoid biasing the cost volume.

Reproducibility artefacts. Release (i) a manifest mapping sequences to environments and calibration parameters, (ii) raw stereo recordings, (iii) per-frame detection logs, and (iv) scripts to recompute metrics. Provide summary spreadsheets mirroring the reporting template for direct comparison.

Scope note. The protocol above describes both standard benchmarking practice (e.g., mAP for multi-class detection) and the practical subset implemented in this thesis. For single-class, single-detector evaluation without ground-truth bounding boxes, classification accuracy and detector confidence serve as tractable proxies. For depth, the per-detection formulation of Valid% and RWB is appropriate when segmentation masks are unavailable—the metrics still capture the key failure modes (matching failure and rear-wall locking) that degrade operator-aid reliability.

Conclusion. This protocol elevates two refractive failure indicators, Valid% and RWB, alongside conventional accuracy metrics. Together they provide quantitative guidance on when stereo ranging of transparent targets is trustworthy and when operator aids must down-weight or suppress range overlays.

Table 2.2: Sensing options for transparent-object perception under water. The second column summarises applicability and extra hardware; the third column gives key cues plus pros/limits.

Modality	**Applicability; extra HW**	**Key cues, pros and limits (underwater transparency)**
RGB **stereo**	**Yes**; none	Boundary/specular cues; multi-view consistency. Pros: simplest, light, real-time; integrates with existing ROV rigs. Limits: fails on index-matched, water-filled targets (rear-wall bias, %NaN); requires refraction-aware calibration/rectification.
RGB **mono + learning**	**Yes (assist)**; none	Context and refractive/edge features. Pros: robust detection; fills gaps when stereo invalid. Limits: depth not physically grounded—must be confidence-gated and range-calibrated.
Polarisation (linear filters)	**Limited**; low	Specular attenuation; weak surface-normal cues. Pros: cheap add-on, can strengthen boundaries. Limits: benefits drop with turbidity/orientation; modest gains on water-filled objects.
Event cameras	**Limited**; medium	High-temporal-resolution edges. Pros: robust to flicker, low latency. Limits: low SNR in low light; underwater algorithms still maturing.
Multispectral/NIR	**Limited**; medium	Material/absorption contrast. Pros: can separate backgrounds in specific conditions. Limits: water absorption shortens range; added complexity.
Light-field (plenoptic)	**Research-only**; high	Angular sampling of refractive flow. Pros: rich cues for refractive surfaces. Limits: bulky, low light efficiency; not fieldable on small ROVs.
Thermal (LWIR)	**No**; medium	Emissivity contrast. Water is opaque in LWIR; not applicable for submerged scenes.
Active range (ToF/structured light)	**Limited**; med–high	Direct depth. Pros: works at short range in clear water. Limits: scattering/attenuation, power/safety constraints; included mainly to justify stereo-first focus.

Table 2.3: Coverage of representative datasets (indicative). "Transp." means transparent consumer containers; "Stereo" means paired, time-synchronised views; "Depth GT" means object-level ground-truth depth for transparent targets.

Dataset	**Underwater**	**Transp.**	**Stereo**	**Depth GT (transp.)**
URPC (2019–2021)	Yes	Limited	No	No
SeaClear (2024)	Yes	Yes	No	No
Trash-ICRA19/UMD	Yes	Yes	No	No
ClearGrasp	No	Yes	Optional	Yes (in air)
Trans10K	No	Yes	No	No
KeyPose/ClearPose	No	Yes	Multi-view	Pose/Depth (in air)

2.9 Summary

This chapter has reviewed the elements that most strongly govern vision-based detection-and-ranging underwater. Image formation (attenuation, backscatter) and refraction through flat ports jointly violate pinhole assumptions, yielding non-central projection and biasing standard stereo. Refraction-aware modelling and calibration—together with strictly matched photometric pre-processing—are therefore prerequisites for usable disparity.

For transparent containers, appearance is dominated by background remapping and specularities; under water, index matching in *water-filled* cases collapses disparity and produces rear-wall estimates and NaNs. In-air methods contribute useful priors (boundary emphasis, multi-view aggregation, confidence gating) but must be adapted to underwater formation and refractive geometry. Among sensing options, RGB stereo remains the most deployable backbone on compact ROVs; polarisation can assist, and monocular learning aids detection but does not replace geometry.

Key takeaways for this thesis:

- Use refraction-aware calibration/rectification (or a validated effective pinhole within a narrow working range); verify residual vertical disparity on submerged planar targets.
- Enforce view-parity in pre-processing and left–right consistency during matching; gate range on confidence to avoid misleading overlays.
- Evaluate not only conventional accuracy (RMSE, Bias, classification accuracy) but also refractive failure indicators: Valid% (success rate) and Rear-Wall Bias (RWB), reported per environment and target type.
- Expect lighting to help *air-filled* transparents but not to solve water-filled, index-matched cases; design experiments and operator aids accordingly.

The next chapter formalises the methodology and calibration/rectification choices adopted here, and the subsequent chapter introduces a multi-environment dataset and protocol that make these failure modes measurable. The MR/AR operator aids later in the thesis consume these perception outputs (range and confidence) rather than redefine them.

Chapter 3

Stereo Vision for Augmented Telepresence and Underwater Object Ranging

3.1 Introduction

In this chapter, the implementation of the 3D stereo vision and mixed reality (MR) teleoperation research project is presented, including the system overview, methodology, and results that guided changes in the setup. The aim is to provide stereo-derived spatial information for MR overlays—primarily distance cues (Fig. 3.1)—to support operator awareness and manipulation in augmented telepresence scenarios [65].

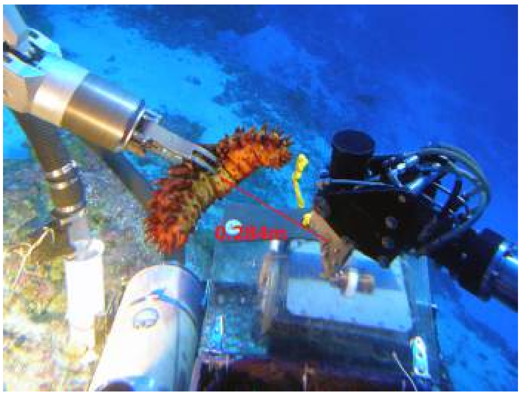


Figure 3.1: Annotated image from an ROV camera of a distance measurement from the manipulator end effector to a sea sponge for collection [65].

Part I: Stereo Vision (Unsynchronised)

3.2 System overview

The stereo vision-based teleoperation interface (Fig. 3.2) comprises a 3D PluraView stereo display to represent video flow from a remote robot's cameras, a desktop haptic interface (Touch), and a custom MR-based robotic manipulator simulation. The aim is a low-cost, safe training environment that combines a virtual robot model with a stereo representation of the remote environment. The 3D PluraView renders stereoscopic images using cross-polarised parallax presentation and a beam-splitter mirror; passive polarised glasses ensure each eye receives the intended view. This provides egocentric depth perception for the remote scene, onto which the virtual end effector is augmented and controlled via the haptic device (stylus position mapped to the end effector in Unity) [33].

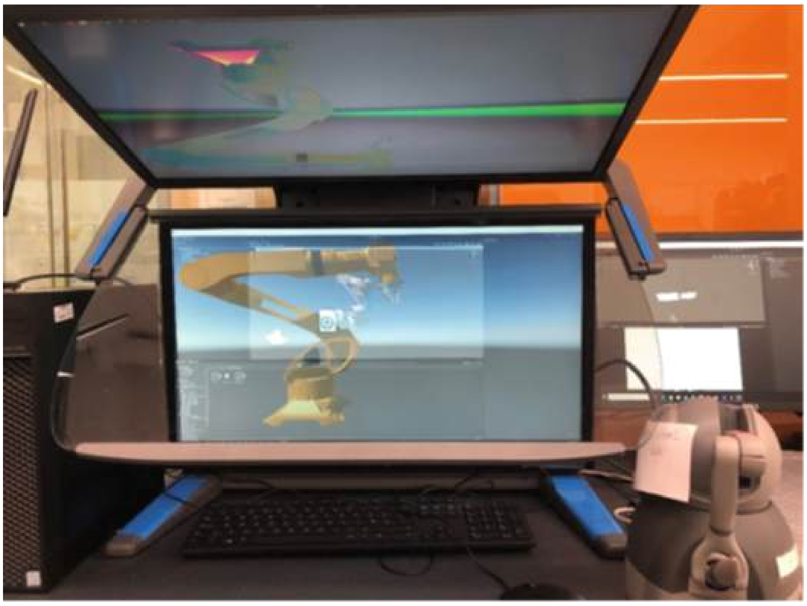


Figure 3.2: 3D PluraView stereo display and Touch haptic controller for teleoperating a virtual robotic end effector. The display uses cross-polarised parallax presentation and a beam-splitter mirror; passive polarised glasses ensure each eye receives the intended view, providing egocentric depth perception for the remote scene.

A high-level flow diagram is shown in Fig. 3.3. Two mono camera streams are rectified and processed in OpenCV, with images and distance data streamed via UDP sockets to Unity. The Unity environment renders a simulated end effector in AR using the Vuforia plugin; a Model Target (generated via the Vuforia Model Target Generator) enables recognition of a real object in the camera stream and triggers a co-located Unity GameObject. The stereo distance to the real object is sent to Unity and subtracted from the virtual end-effector to virtual-object distance so the overlayed measurement remains consistent.

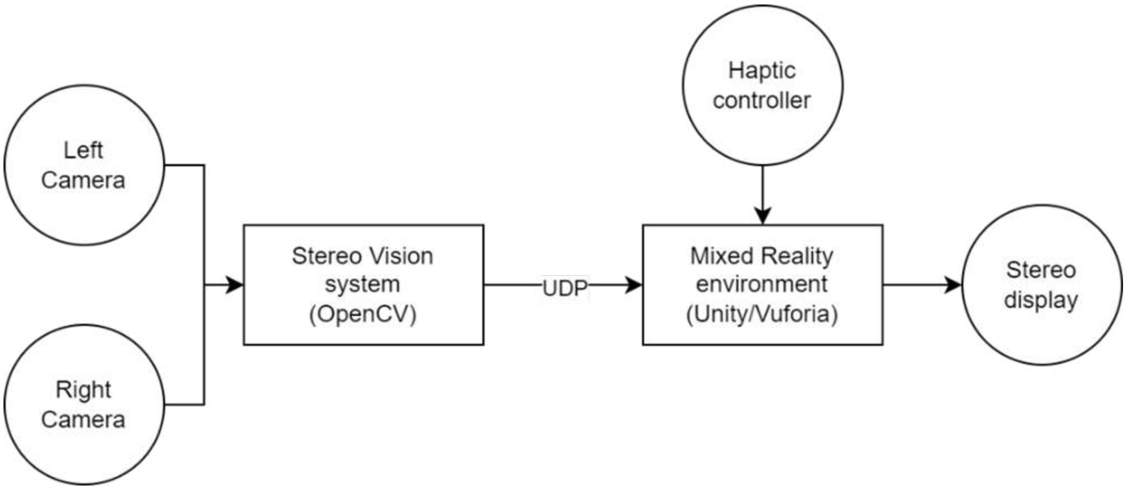

Left Camera
Right Camera
Stereo Vision system (OpenCV)
UDP
Haptic controller
Mixed Reality environment (Unity/Vuforia)
Stereo display

Figure 3.3: High-level unsynchronised system diagram: OpenCV stereo node produces disparity, depth and distance estimates; Unity renders MR overlays; data flow is via UDP sockets.

3.3 Stereo vision methodology

Two webcams (Logitech C505) were mounted with parallel optical axes and a horizontal baseline close to the human interpupillary distance (IPD, average 63 mm), which also falls within the stereo zone of comfort [71]. The processing pipeline is summarised in Fig. 3.4 and includes calibration, rectification, dense matching, and distance estimation following standard formulations [1, 33, 77].

Calibration and rectification. Checkerboard images were captured at varied poses to estimate intrinsics (focal length, principal point) for each camera and the stereo extrinsics. Rectification removes lens distortion and reprojects the two images onto a common plane so correspondences lie along horizontal epipolar lines [1, 33].

Dense correspondence. OpenCV's `StereoBM` was used to generate disparity maps. `Stereo BM` is a real-time dense method (1920 × 1080 in milliseconds) well-suited to textured scenes; `Stereo-SGBM` provides more robust sub-pixel matching on smoother textures at the expense of speed [33, 60]. Block matching follows three steps [69, 77]:

1. **Prefilter**: normalise brightness/enhance texture with a sliding window (e.g., 5×5 to 21× 21).
2. **Search**: compute similarity (SAD) along epipolar lines over a disparity range; window size and disparity bounds strongly affect runtime/accuracy.

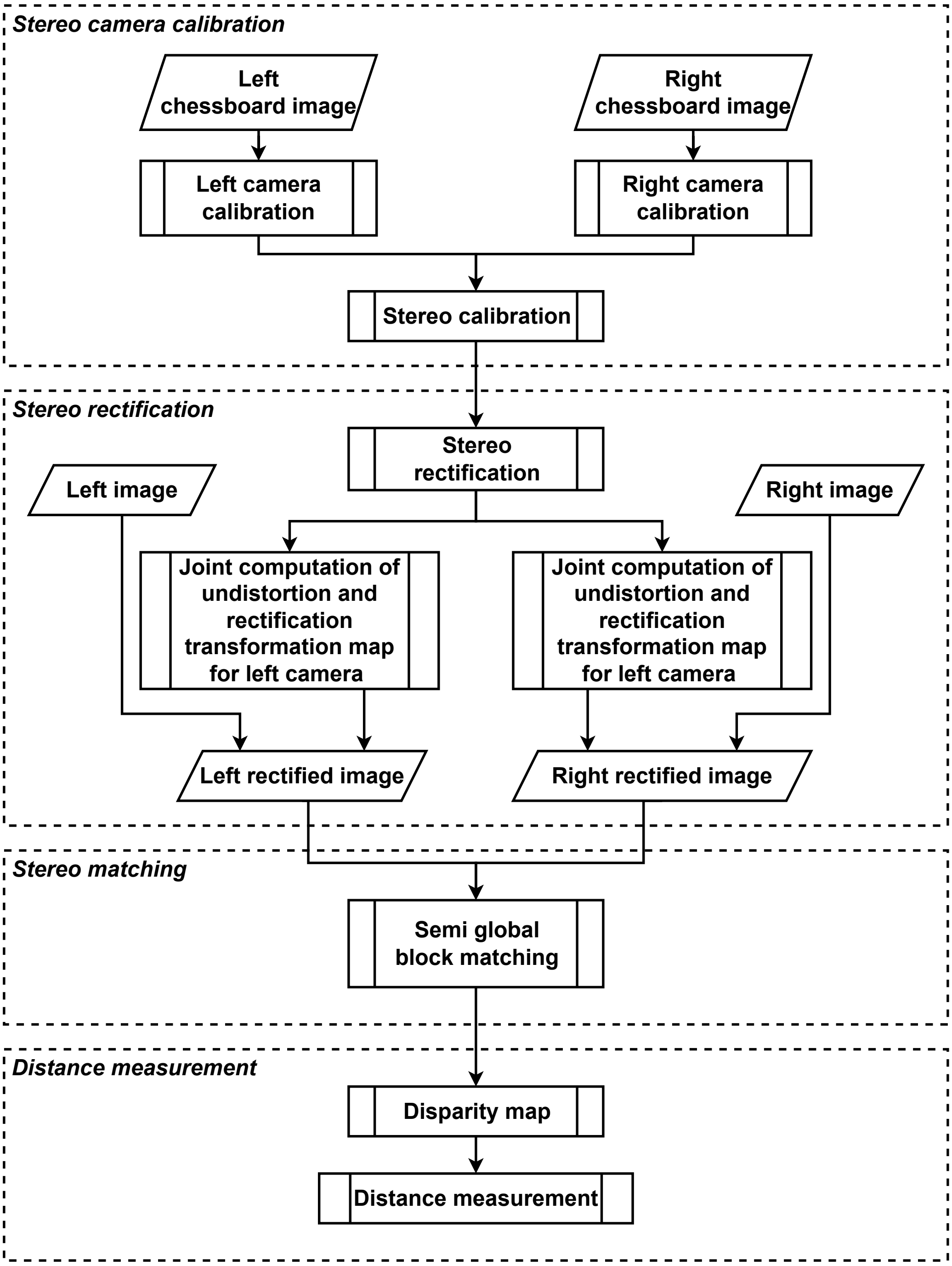


Figure 3.4: Flow diagram of the proposed stereo vision distance measuring system: calibration, rectification, dense matching, and distance estimation [33, 77].

3. **Best match & postfilter**: select the minimum cost and reject poor matches via heuristics such as `uniquenessRatio`; reduce small isolated mismatches with `speckleWindowSize`.

Fig. 3.5 (single image with two panels) illustrates the scanline SAD process and a typical disparity output.

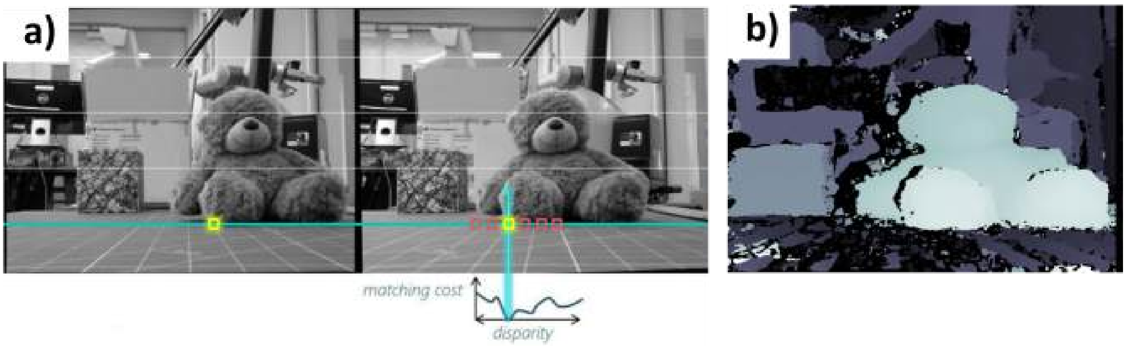


Figure 3.5: Dense stereo block matching and disparity example: (a) scanline and candidate matches with SAD cost; (b) disparity image computed in OpenCV.

Distance estimation. Triangulation from disparity d yields depth z via

$$z = \frac{f\,b}{d},$$

with f the focal length in pixels and b the baseline [1, 33]. For MR overlays, distances were computed within a region-of-interest (ROI) mask and summarised with robust statistics to stabilise readouts.

3.4 Validation and limitations

Cameras were calibrated with reprojection errors of 0.0131 px for both left and right, at a 6.3 cm baseline. Range testing used a tape measure for ground truth across approximately 45–130 cm, with examples of disparity shown in Fig. 3.5 (panel b). The object set comprised a patterned cardboard box, metal bottle, clear bottle, white bottle, a fabric teddy toy, and a rough-textured 3D-printed part. Eighteen distances were tested (ten readings each). Results are summarised in Fig. 3.6, which is a single composite image containing both (a) the objects and (b) the distance estimates. Per-object RMSEs were: box 14.57 cm; clear bottle 16.34 cm; white bottle 1.01 cm; metal bottle 2.03 cm; teddy 1.95 cm; 3D-printed part 2.50 cm. The average RMSE was 9.13 cm. Clear-bottle readings occasionally returned 0 cm (null) due to insufficient matches, but were accurate when valid disparities were present. In practice, masking the object ROI and averaging within-mask points mitigates outliers from disparity speckle and holes [1, 33].

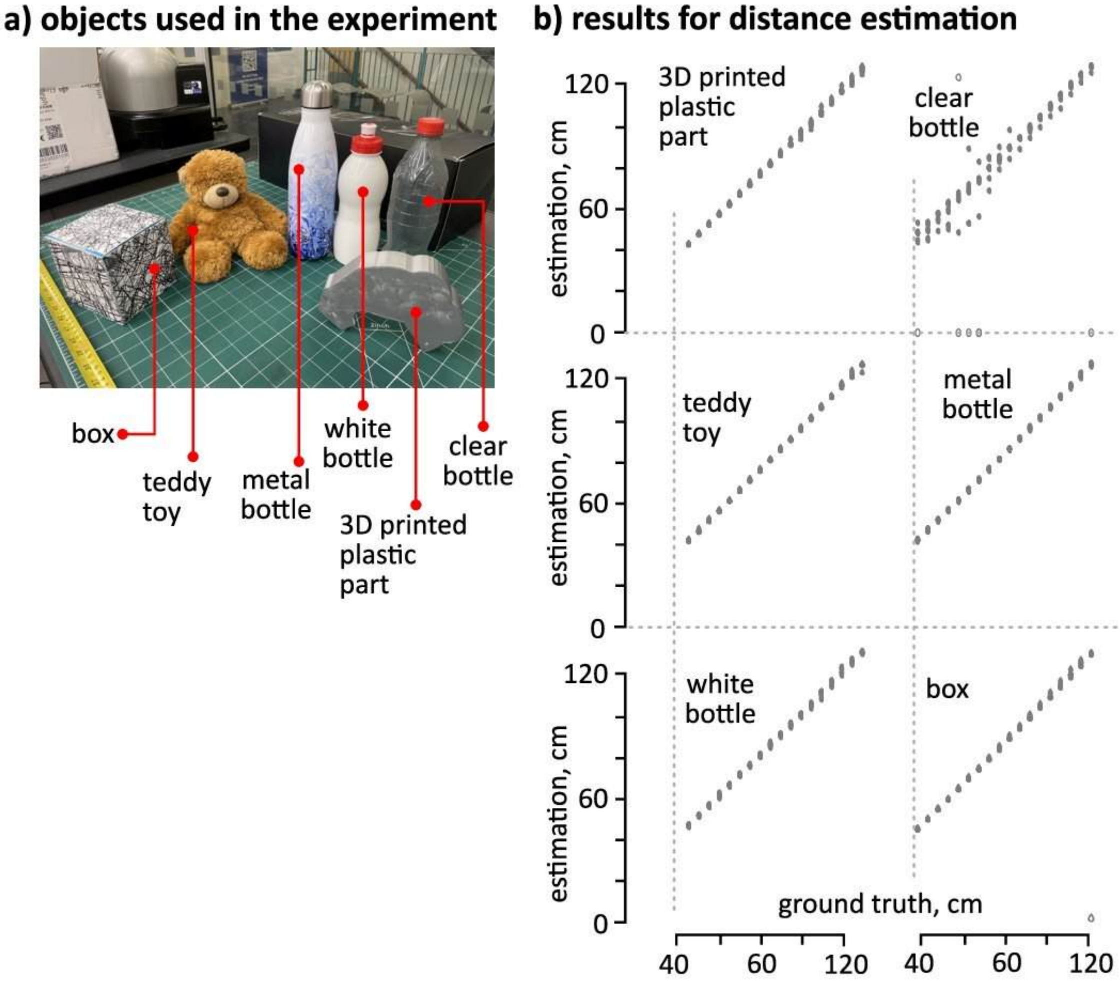


Figure 3.6: Unsynchronised validation in air (single composite image): (a) object set; (b) distance estimation results across ∼45–130 cm (ten readings per distance).

Whilst the unsynchronised system validated stereo ranging in air, teleoperation validation was blocked by critical incompatibilities: asynchronous capture vs rendering, transport jitter, and the AR plug-in's monocular constraint (Fig. 3.7). These issues motivated a synchronised stereo front-end with a unified SDK.

Part II: Synchronised Stereo Vision and Underwater Object Ranging

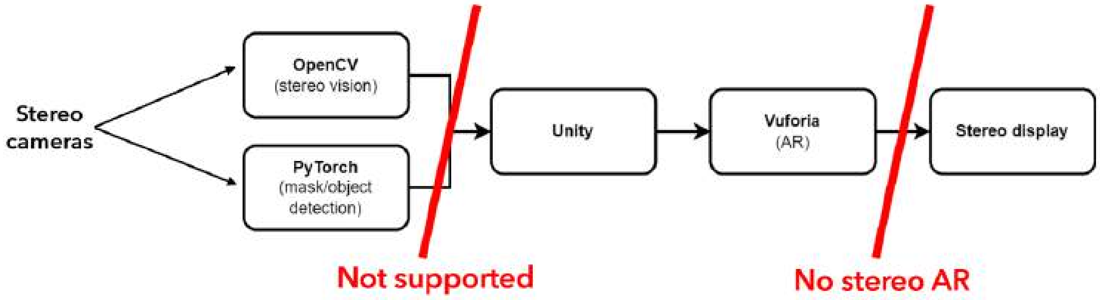


Figure 3.7: Critical blockers in the unsynchronised pipeline across capture, transport, and AR rendering.

The unsynchronised pipeline demonstrated that commodity stereo can deliver useful distance cues in air, but also exposed fundamental blockers (Fig. 3.7)—asynchronous capture, transport jitter and AR plug-in constraints—that prevented stable MR overlays during teleoperation. To address these limitations, this part transitions to a synchronised stereo camera with a vendor-supplied SDK that provides hardware-timestamped stereo pairs, on-device depth, and native Unity integration. This eliminates the capture–render synchronisation problem and enables real-time MR overlays that remain stable under operator motion. Building on this more robust foundation, we then extend the stereo rig to an underwater testbed, recalibrate behind a Perspex window, and run systematic experiments to characterise how medium properties and target materials affect ranging performance for transparent bottles.

3.5 System overview

Using a ZED Mini stereo camera (ZEDM) enabled us to bypass the previous blockers by replacing OpenCV/PyTorch and Vuforia components with the ZED SDK (Fig. 3.8). The Unity application ingests vendor-provided stereo/depth and pose, stabilising the MR overlays under operator motion.

3.6 Stereo-informed augmented telepresence

A 3D manipulator model (URDF) is imported into Unity using the XR Toolkit and controlled via a Geomagic Touch[1] haptic device mapped through ROS#[2]. The camera object is registered

[1] https://www.3dsystems.com/haptics-devices/touch

[2] https://www.ros.org/

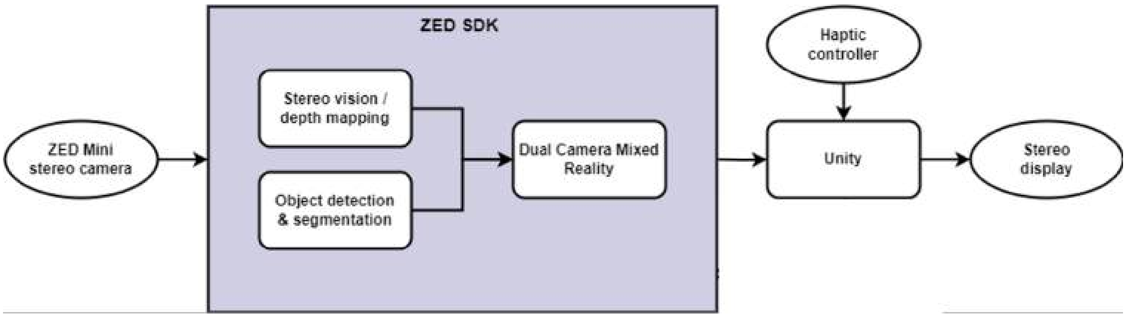

ZED SDK
ZED Mini stereo camera
Stereo vision / depth mapping
Object detection & segmentation
Dual Camera Mixed Reality
Haptic controller
Unity
Stereo display

Figure 3.8: System diagram using the ZED Mini camera and ZED SDK integrated with Unity.

in Unity so the distance between the simulated end effector and a physical object is calculated by subtracting the stereo-measured object distance. In air, the manufacturer specifies $\lesssim 15\,\text{mm}$ error at 1 m; this was validated as a baseline prior to underwater recalibration. Distance was also sonified: a thresholded bleep frequency increased as the end effector approached the object.

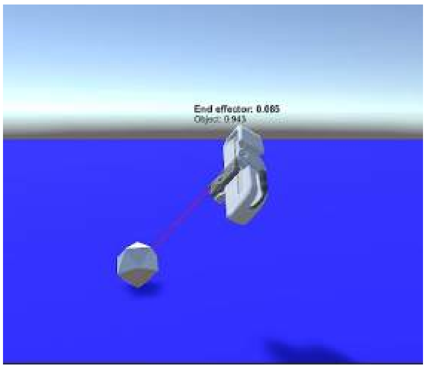


Figure 3.9: Unity model of the simulated end effector with laser-pointer distance cue in virtual space (and mirroring in MR scenes).

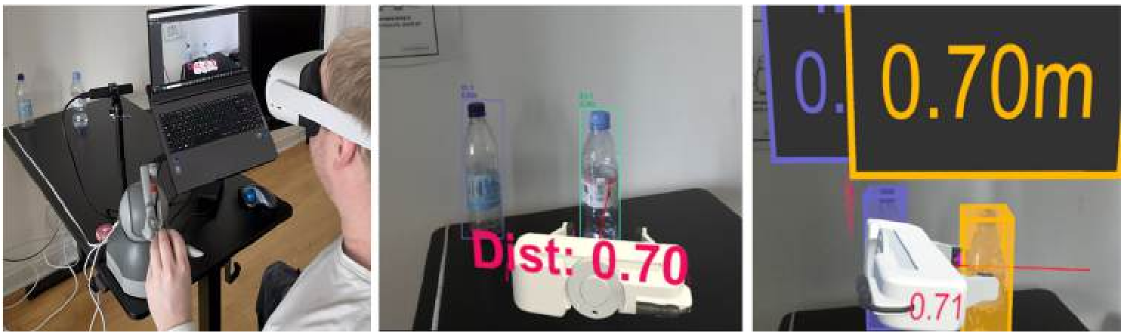
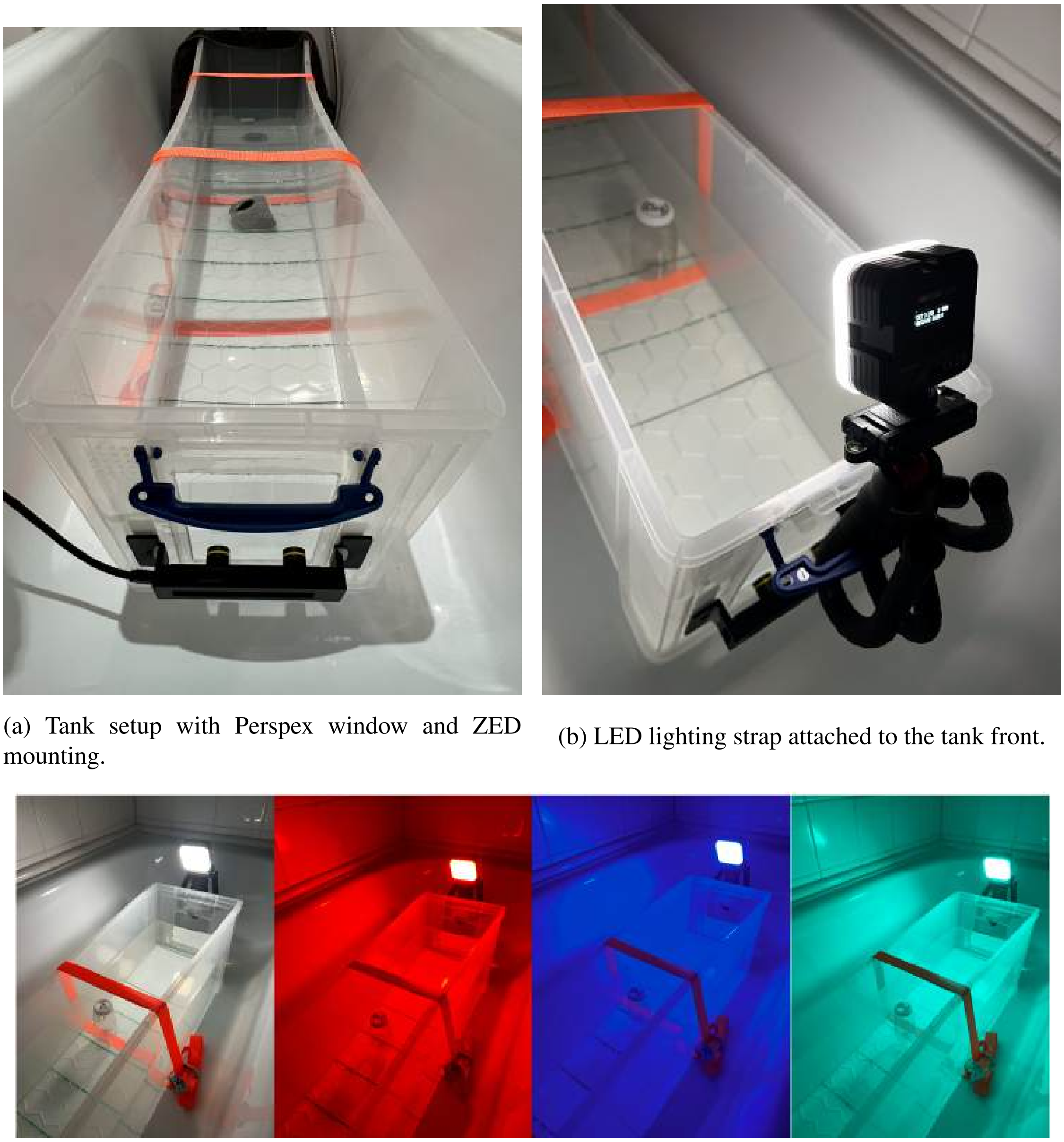


Figure 3.10: Mixed-reality headset view (VR HMD) showing augmented distance cues and a virtual end effector, controlled via a haptic controller.

3.7 Experimental setup and methods

Since the ZEDM is not waterproof, an experimental tank was constructed. A plastic storage box was modified with a clear Perspex window, sealed with aquarium sealant; a custom 3D-printed mount attached the ZEDM to the window. An LED multicolour light strap was mounted to the front to provide directional and coloured lighting (Fig. 3.11). Recalibration was performed for underwater use, following the vendor's approach; a reprojection error of 0.22 px was achieved.



(a) Tank setup with Perspex window and ZED mounting.

(b) LED lighting strap attached to the tank front.

(c) Lighting conditions reference panel (WRGB and Ambient scenes).

Figure 3.11: Experimental tank setup with ZED stereo camera mounted to a Perspex panel window (a) and a LED multicolour light strap to front pointing into the tank (b) with different lighting conditions being tested (c)

Distance/medium experiment. Six material types were tested: white bottle, stone, textured clear bottle, smooth clear bottle, filled textured clear bottle, and filled smooth clear bottle. Three environments were used: air, underwater without recalibration, and underwater with recalibration. Ground-truth distances spanned 300–1000 cm in 100 cm steps (air limited to 700 cm). For each material, distance, and environment, we acquired 10 repetitions (total ∼480 measurements across all conditions). The underwater object set is shown in Fig. 3.12.

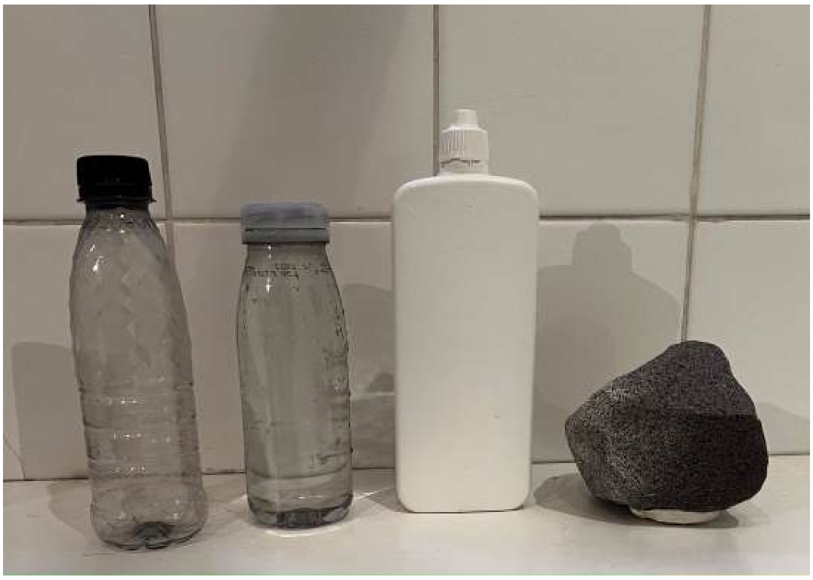


Figure 3.12: Objects used in the underwater experiments: textured clear bottle, smooth clear bottle, white bottle, and stone (additional filled transparent variants appear in the results).

Lighting experiment (air). A complementary study assessed the smooth clear bottle under six light colours (Ambient, White, Red, Blue, Green, Yellow) across the same distances (300–1000 cm; 100 cm steps), with 10 repetitions per colour per distance (480 measurements).

RMSE aggregation. To aggregate per-distance errors consistently across ranges, we report the overall RMSE as the root-mean-square of per-distance RMSEs:

$$\mathrm{RMSE}_{\mathrm{overall}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{RMSE}_i^2}.$$

3.8 Results and discussion

3.8.1 Materials × environment

Figure 3.13 shows per-material distance estimates in air, underwater uncalibrated, and underwater recalibrated conditions. In air, all materials except the smooth clear bottle exhibited excellent linearity and low variance up to 700 cm. Underwater without recalibration, distances were consistently underestimated (systematic bias), especially at longer ranges. Recalibration substantially reduced bias for opaque/textured materials, with filled transparent bottles remaining the most variable. The white bottle and stone were the most consistent across environments.

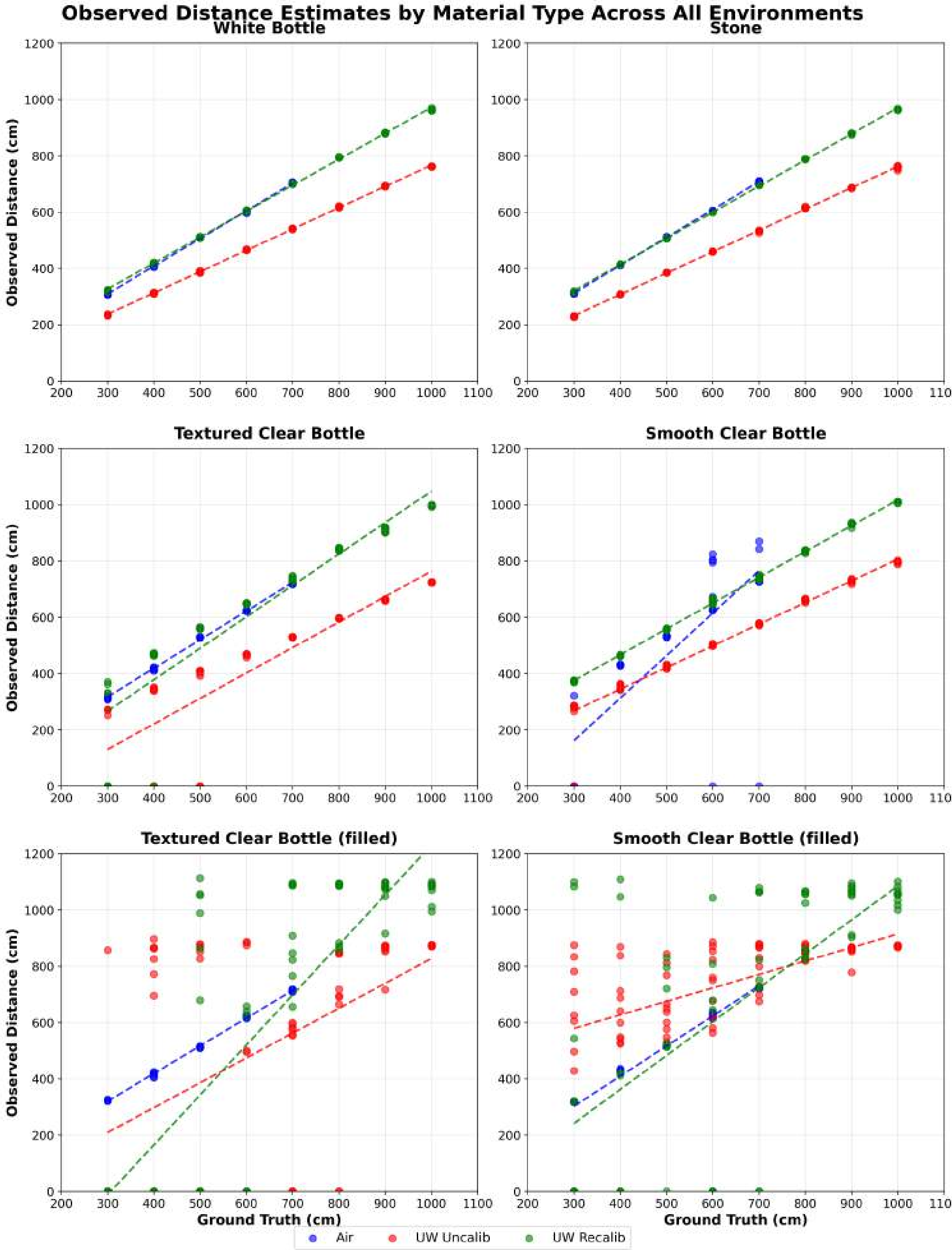


Figure 3.13: Observed distance by material across environments (air, underwater uncalibrated, underwater recalibrated). Trend lines indicate linear fits per environment.

Figure 3.14 summarises RMSE across materials and environments. In air, white bottle (8.8 cm), stone (10.7 cm), and textured clear bottle (20.2 cm) performed best; the smooth clear bottle showed anomalous behaviour in air (194.7 cm). Underwater uncalibrated conditions degraded performance broadly (e.g., filled textured clear bottle 383.0 cm). Recalibration improved opaque materials markedly (white bottle 18.5 cm; stone 17.4 cm), but filled transparent bottles remained high-error (327–340 cm), indicating fundamental challenges when the object is water-filled and submerged.

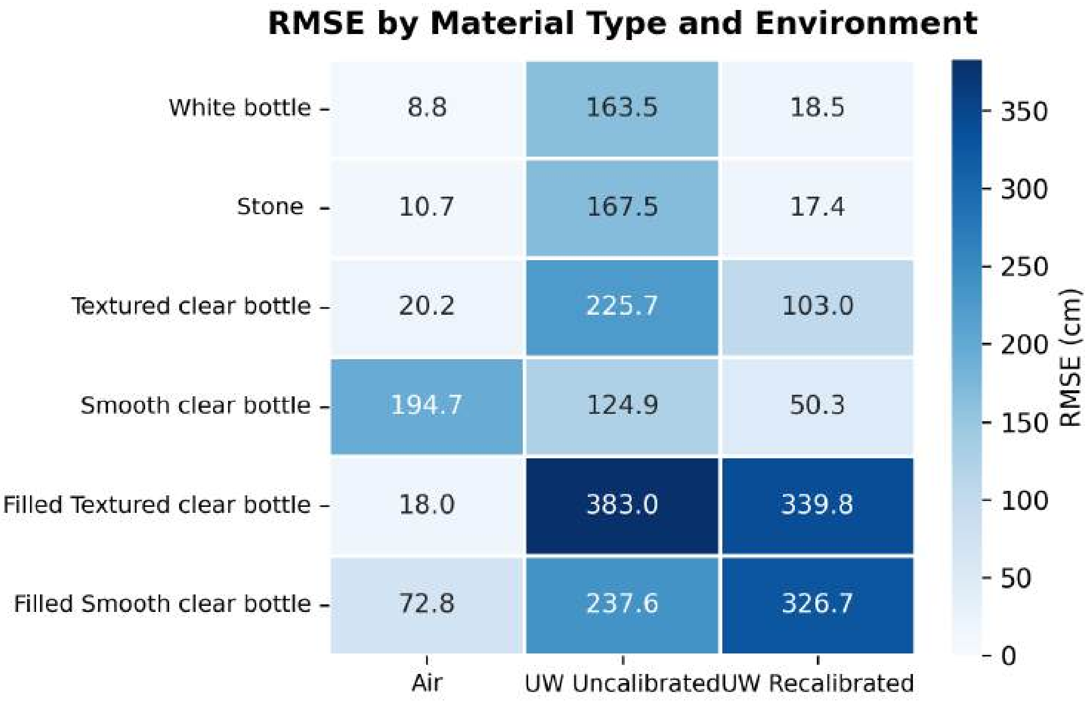


Figure 3.14: RMSE heatmap by material and environment. Recalibration closes the gap for opaque/textured materials but not for water-filled transparent bottles.

3.8.2 Lighting effects (air)

Figure 3.15 shows distance estimates under different light colours. Most colours (Ambient, White, Red, Green, Yellow) clustered close to the ideal trend; white light performed best. Blue light was a severe outlier, with an inverted trend and very large errors.

The quantitative analysis (Fig. 3.16) confirms white light as best (RMSE 35.4 cm); red and green were 40.1 cm, yellow 43.3 cm, ambient 50.3 cm. Blue light failed catastrophically (585.3 cm), suggesting a mismatch between narrow-band blue illumination and reliable stereo correspondence on this object at range.

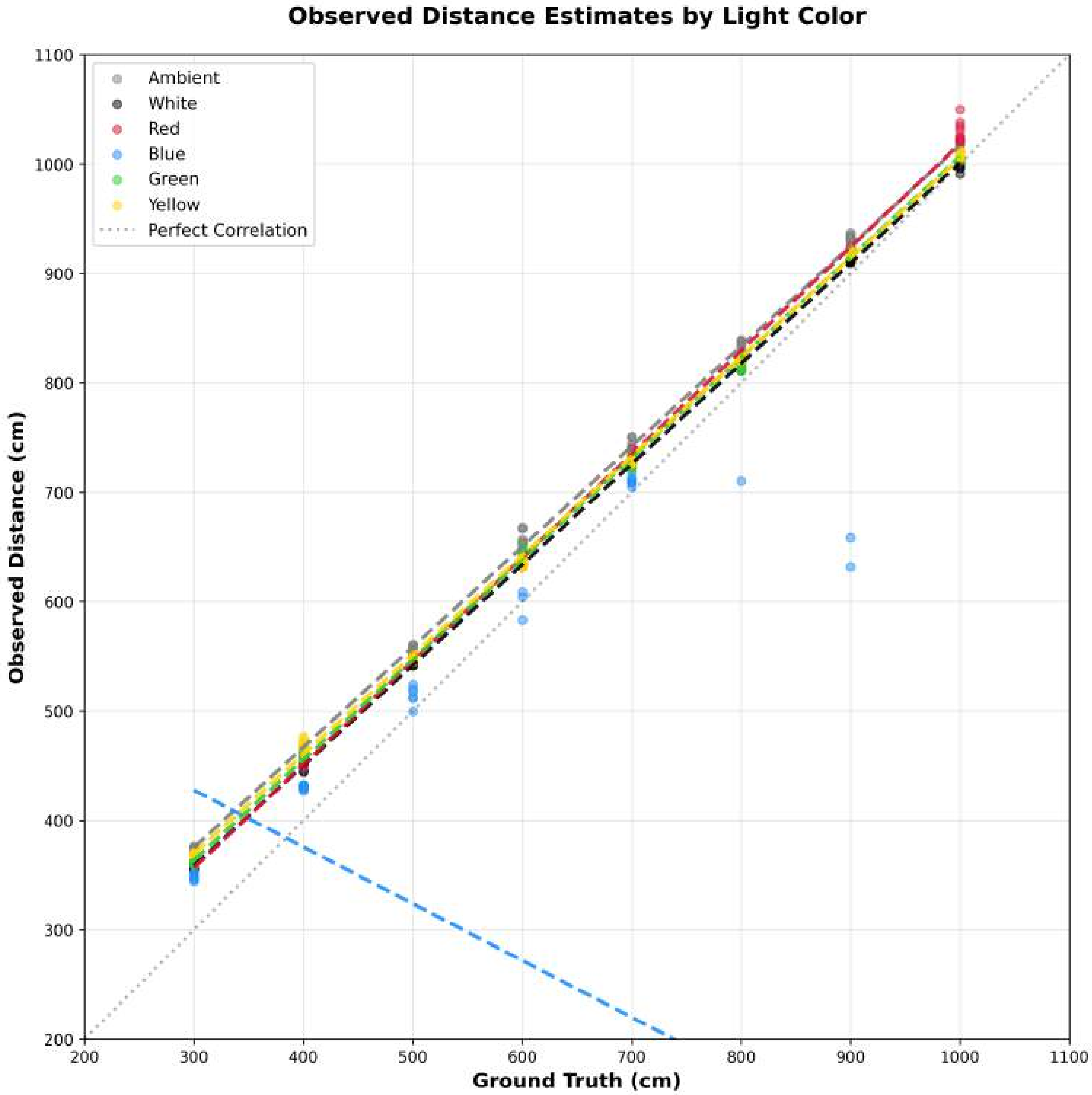


Figure 3.15: Observed distance under different light colours (Ambient, White, Red, Blue, Green, Yellow) for the smooth clear bottle in air.

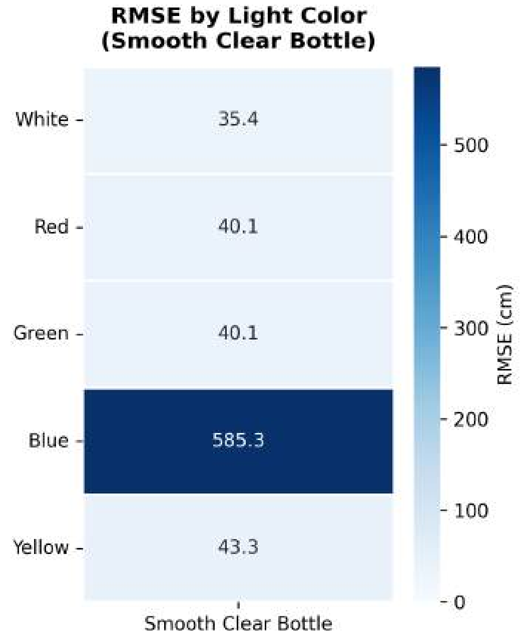


Figure 3.16: RMSE by light colour for the smooth clear bottle in air. White is best; blue causes failure across the range.

3.8.3 Implications

Synchronising the stereo front-end and integrating the vendor SDK improved the stability and interpretability of MR distance cues, particularly for opaque/textured targets and under controlled lighting. Underwater, recalibration was essential but insufficient for water-filled transparent plastics: when a smooth clear bottle is air-filled, the depth engine returns a surface; when it is water-filled and submerged, the bottle often appears effectively see-through, with the system reporting the next visible surface (e.g., the tank back wall). Figure 3.17 illustrates this phenomenon. These findings motivate the next chapter's focus on underwater transparent-object detection and ranging.

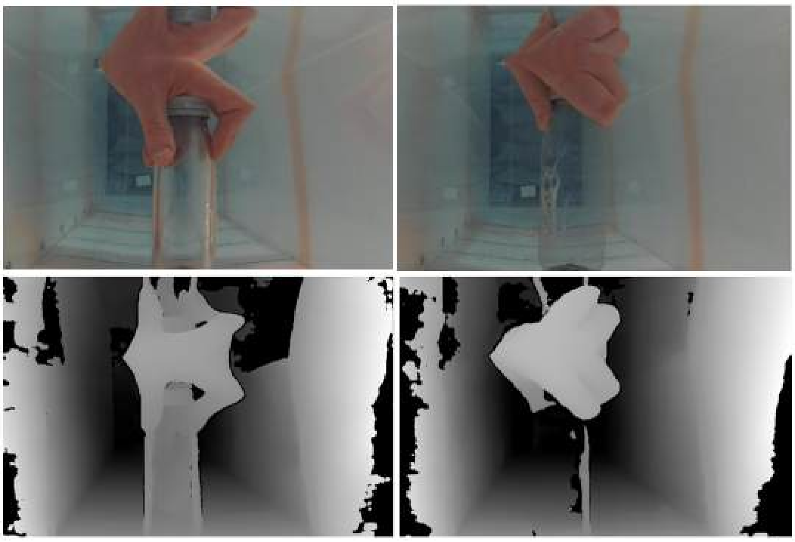


Figure 3.17: Transparent bottle behaviour underwater. Top: RGB; bottom: depth map. Left: air-filled bottle (valid surface). Right: water-filled bottle (see-through in depth).

3.9 Summary

In this chapter, we implemented an unsynchronised OpenCV → Unity stereo pipeline for MR teleoperation, validated its distance estimation in air, and analysed its limitations, motivating a transition to a synchronised ZED SDK-based pipeline [33, 77, 1]. We then built an underwater testbed with a recalibrated stereo headset camera and WRGB lighting, introduced a materials × environment protocol and a controlled lighting study, and reported accuracy trends across conditions. Recalibration improved opaque and textured materials underwater, while water-filled transparent bottles remained unreliable, especially under blue light. These results inform the design of stereo-informed augmented telepresence and establish the need for dedicated methods for underwater transparent-object detection and ranging in the following chapter.

Chapter 4

UW-TransStereo: a multi-environment underwater stereo dataset for transparent bottle detection and ranging

4.1 Introduction & contributions

Transparent and refractive objects pose well-known challenges for passive stereo vision: specularities, refraction at curved interfaces, low contrast due to attenuation/backscatter, and correspondence ambiguity degrade both detection and depth estimation. This chapter consolidates and expands two of our publications to introduce *UW-TransStereo*, a stereo perception dataset and benchmark focused on plastic bottle detection and ranging across air and multiple underwater media, together with a reproducible capture-and-analysis pipeline [74, 74, 75].

Scope and rationale. We target end-to-end *detection-and-ranging* with a ZED stereo rig and a modern object detector (YOLOv8), quantifying performance across six test conditions: Air, Freshwater Day, Freshwater Night, Freshwater with an inference-time enhancement ("basicEnhance"), Saltwater, and Saltwater with suspended Pellets [74]. For ranging we report both *manual point distance* (mean of three vertical samples within each detection box) and the SDK's *automatic object distance*, enabling like-for-like comparison of per-pixel and per-object estimates [74]. The expanded analysis here generalises the freshwater-focused subset previously reported and released, and introduces a cross-environment summary via RMSE heatmaps for distance and accompanying accuracy/confidence panels [74, 74].

Contributions. This chapter makes the following contributions:

1. **Dataset and release.** We introduce *UW-TransStereo*, a multi-environment stereo dataset for transparent bottle detection & ranging [74]. Each sequence pairs a raw rectified stereo recording (`.svo2`) with per-frame detection logs (`.xlsx`) and qualitative screenshots, organised by environment and bottle type; collated CSVs expose fields for detector confidence, three sampled point distances (upper/centre/lower), object distance, 3D size, and timestamps [74]. This mirrors the format used throughout our analysis and is designed for reproducible benchmarking [74].

2. **Reproducible capture & logging pipeline.** We provide an end-to-end pipeline that integrates SVO recording and deterministic logging (automatic file naming, structured storage of videos, spreadsheets, and screenshots), with real-time visualisations (2D boxes, 3D point cloud, bird's-eye tracking) and runtime controls (pause/resume, CLI-configurable model and thresholds). The tooling supports replay-based evaluation and exact condition matching across experiments [74].

3. **Six-condition benchmark with dual range estimators.** We benchmark detection accuracy, detector confidence, and ranging error (RMSE) for both manual point distance and automatic object distance across six environments (Air $\rightarrow$ Saltwater+Pellets), and provide a compact cross-environment heatmap that aligns trends across metrics and bottle types [74]. This enables diagnosis of dominant failure modes (e.g., filled smooth transparent bottles) and the environment-dependent degradation of success rates and error [74].

4. **Statistical significance analysis (Saltwater vs Freshwater).** We include a formal comparison showing that saltwater yields significantly lower distance error than freshwater for both average point distance and object distance (MAE improvements of 58.7 cm and 31.2 cm, respectively; $p < 0.001$ with small but practically meaningful effect sizes), higher object-distance success rates (47.4% vs 41.6%, $p = 0.023$), and no significant difference in classifier-score distributions ($p = 0.598$) (see Section 5.1.2).

5. **Practical guidance for deployment.** We evaluate lightweight inference-time strategies (success-only filtering, confidence gating, temporal smoothing) and an illustrative preprocessing variant ("basicEnhance"). These stabilise detection presence and can improve success rates, but introduce a stability–precision trade-off and do not close the air–water gap for refractive cases; environment- and range-aware policies (e.g., 55–60% confidence

thresholds in Freshwater Day, higher in more challenging conditions) are recommended [74].

Positioning. In the broader context of transparent-object perception and underwater imaging, UW-TransStereo complements existing in-air datasets and algorithmic advances by focusing on refractive targets underwater with stereo, delivering paired raw data and structured logs suitable for both classical and learning-based approaches [74]. The remainder of the chapter details the capture setup and data organisation (§2), the processing and evaluation methodology (§3), and the reproducibility resources (§4). The comprehensive multi-environment results analysis with cross-environment heatmaps and per-condition plots is presented in Chapter 5.

4.2 Related context

Stereo perception of transparent and refractive objects is challenging both in air and underwater, where specularities, refraction at curved interfaces, and backscatter degrade correspondence and depth estimation. Recent in-air work on geometry-aware depth completion and pose estimation for transparent objects (e.g., ClearGrasp and GDR-Net) demonstrates that combining learned priors with multi-view constraints can recover object geometry despite missing or distorted depth [74, 67, 86]. Our multi-environment study is designed to probe how similar ideas might extend to underwater refractive cases, where light propagates through layered media and refractive index mismatches are stronger [74].

The findings reported in Chapter 5 motivate three broad research directions: (i) geometry-aware methods that explicitly model object shape and refraction; (ii) physics-based correction, for example via ray-tracing through known refractive interfaces or learned refraction fields; and (iii) multi-modal fusion that combines stereo with complementary modalities such as polarisation imaging, structured light, or time-of-flight ranging [74]. UW-TransStereo is structured to support such developments by pairing raw rectified stereo recordings (`.svo2`) with per-frame detector outputs and ground-truth ranges across six environments, enabling both classical stereo analysis and learning-based approaches [74, 74].

4.3 System and experimental setup

4.3.1 Experimental tank, camera, and targets

We used a bench-top tank (approx. H30×W26×D116 cm) with marker lines on the base at 10 cm intervals to standardise object placement and a blue cloth attached to the rear wall. The water fill line was set at 18 cm, yielding an operating volume of approximately 54 litres; for saltwater conditions, 1.9 kg of sea salt was dissolved to approximate average ocean salinity (∼35 g/L). A ZED Mini stereo camera was mounted to a Perspex viewing window cut into the front of the tank [74] (Fig. 4.1). The bottle set comprises five targets representative of transparent debris with an opaque control bottle: *medicine*, *smooth (empty)*, *smooth (filled)*, *textured (empty)* and *textured (filled)* [74] (Fig. 4.2). This enables us to investigate how surface texture and fill state affect visibility (detection and ranging) across multiple environments (air, freshwater, saltwater, saltwater with pellets) (Fig. 4.3) [74].

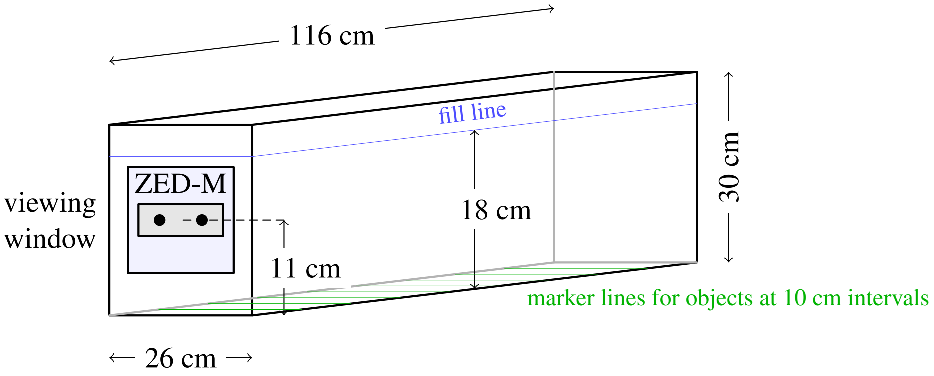


Figure 4.1: Experimental tank schematic showing the perspex viewing window, ZED Mini mounting, water fill line, and equidistant 10 cm placement markers on the tank base [74].

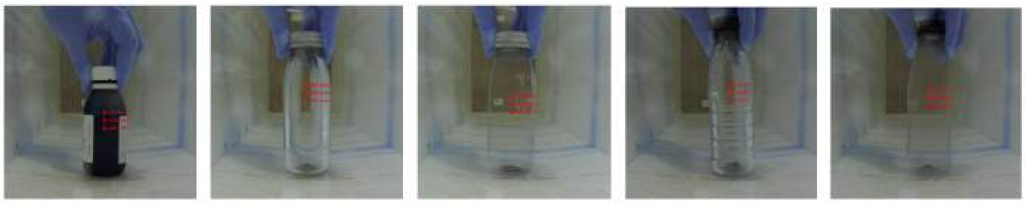


Figure 4.2: Bottle set used in all recordings (left to right): medicine, smooth, smooth filled, textured, textured filled [74].

4.3.2 Acquisition pipeline and runtime controls

The end-to-end pipeline performs stereo calibration and depth, optional edge/depth-aware confidence preprocessing, YOLOv8 detection, ZED SDK ingestion/tracking, and per-detection log-

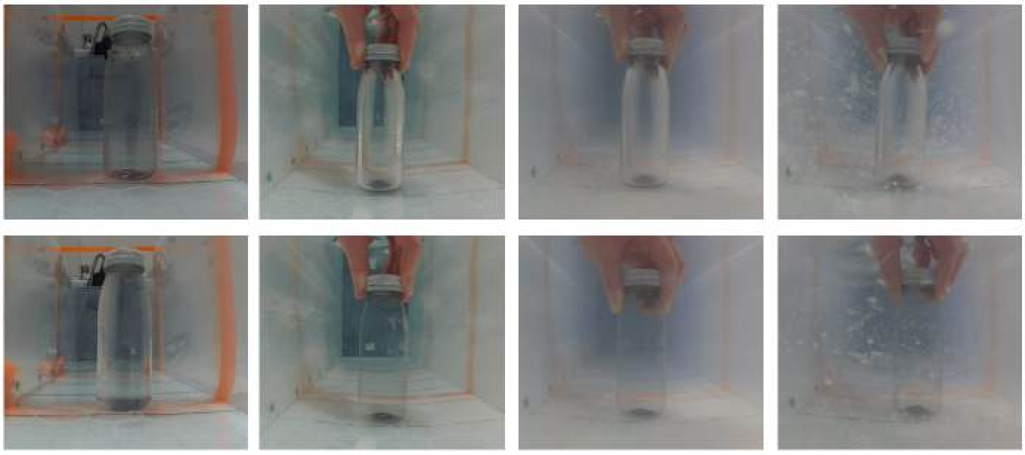


Figure 4.3: Smooth clear bottle air-filled (top) and water-filled (bottom), shown across the multi-environment set (left to right): air, freshwater, saltwater, saltwater with pellets [74].

ging (Excel, screenshots, SVO, configuration) [74]. For reproducibility and experiment control, the capture tool provides SVO recording/replay, automatic timestamped file naming, a structured on-disk layout (videos, spreadsheets, screenshots), real-time 2D/3D/bird's-eye visualisation, pause/resume, and CLI switches for model weights, thresholds, and feature toggles [74].

We also expose a minimal inference-time enhancement (*basicEnhance*) that linearly down-weights confidence with range and blends Canny edges with RGB; parameters are fixed for all runs to preserve comparability [74]:

$$\text{conf}' = \text{conf} \cdot \frac{1}{1 + d_{\text{mm}}/1000},$$

where d_{mm} is the centre depth in millimetres [74].

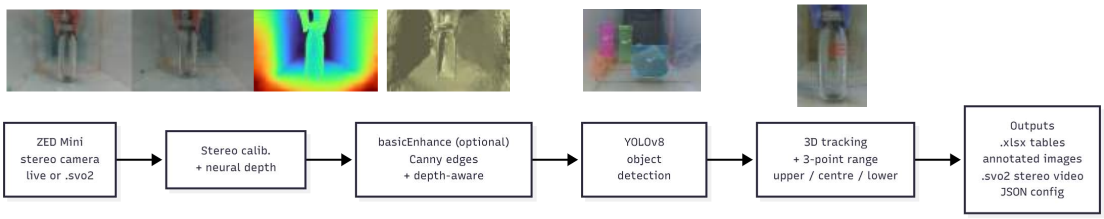


Figure 4.4: Processing pipeline with example screenshots: stereo calibration/depth → optional edge/depth-aware confidence → YOLOv8 detection → ZED ingestion/tracking → per-detection measurements and logs [74].

4.4 Dataset description and records

4.4.1 File modalities, naming, and archive layout

Each recording is released as a paired raw stereo video and per-frame detections, with qualitative screenshots for audit. Specifically: `.svo2` (HD2K, 15 FPS, lossless), one `.xlsx` per recording (class, confidence, box), and `.png` screenshots [74]. Filenames are flat-structured and encode context as `[Environment]_[BottleType]_[filekind]_[timestamp].[ext]` (e.g., `Saltwater-pellets_smtBtl-fill_detections_08052025_183311.xlsx`); timestamps align videos with their spreadsheets and screenshots [74].

The public archive comprises 25 paired recordings (50 core files) across Air, Freshwater, Freshwater-basicEnhance, Saltwater, and Saltwater-pellets, plus 216 screenshots; collated analysis tables exceed 5,000 labelled measurements [74]. A note on time-of-day: the released freshwater raw videos correspond to *night* sequences; day sequences used in earlier comparisons were not recorded to `.svo2` and are therefore not in the archive [74].

Table 4.1: Archive manifest (excerpt). Path conventions and file purposes [74].

Path/Filename (relative)	Type	Purpose
`Air/medBtl/recording_07052025_234742.svo2`	SVO2	Raw rectified stereo
`Air/medBtl/detections_07052025_234742.xlsx`	XLSX	Per-frame detections
`Air/medBtl/screenshot_*.png`	PNG	Qualitative screenshots
`Freshwater/medBtl/recording_08052025_020841.svo2`	SVO2	Raw rectified stereo
`Freshwater/medBtl/detections_08052025_020841.xlsx`	XLSX	Per-frame detections
`Freshwater/medBtl/screenshot_*.png`	PNG	Qualitative screenshots
`Saltwater/medBtl/recording_08052025_182447.svo2`	SVO2	Raw rectified stereo
`Saltwater/medBtl/detections_08052025_182447.xlsx`	XLSX	Per-frame detections
`Saltwater/medBtl/screenshot_*.png`	PNG	Qualitative screenshots
`Saltwater_pellets/medBtl/recording_08052025_191[illegible].svo2`	SVO2	Raw rectified stereo
`Saltwater_pellets/medBtl/detections_08052025_19[illegible].xlsx`	XLSX	Per-frame detections
`Saltwater_pellets/medBtl/screenshot_*.png`	PNG	Qualitative screenshots

4.4.2 Data records (schema)

Per-detection records include ground-truth range, stereo-derived distances sampled at three vertical ROIs (upper/centre/lower), detector confidence, environment/bottle identifiers, object distance (SDK estimate), 3D size, and timestamps; missing values are preserved as NaN and units documented per column [74]. A dictionary excerpt is reproduced in Table 4.2 [74].

Table 4.2: Selected columns in collated CSVs (dictionary excerpt) [74].

Column	Unit	Description / Policy
`environment`	–	{Air, Freshwater, Saltwater, Saltwater_pellets}
`bottle_type`	–	{medBtl, smtBtl, texBtl} with optional `-fill` suffix
`class`	–	COCO class ID (39 = bottle)
`name`	–	Detection class name
`confidence`	[0,1]	Detector score (parsed from percentage)
`center_distance`	cm	Stereo range at centre ROI; NaN if unranged
`upper_distance`	cm	Stereo range at upper ROI; NaN if unranged
`lower_distance`	cm	Stereo range at lower ROI; NaN if unranged
`object_distance`	cm	SDK per-object range
`width`, `height`, `depth`	cm	3D object size
`GT_*`	–	Ground truth fields
`timestamp`	–	ISO time reference

Object dimension fields. In addition to range estimates, the dataset records per-detection 3D object dimensions (`width`, `height`, `depth`) as computed by the ZED SDK's bounding-box estimator. Ground-truth dimensions (`GT_Width`, `GT_Height`, `GT_Depth`) are included for the five bottle types (Table 4.3). These fields enable evaluation of the stereo pipeline's object sizing capability in addition to ranging—a distinct but related task that faces similar challenges from refraction and low contrast. Analysis of dimension accuracy is presented in Section 5.1.7 and Appendix B.

Table 4.3: Ground-truth dimensions for the five bottle types used in all recordings. Dimensions measured manually in millimetres.

Bottle Type	GT Width (mm)	GT Height (mm)	GT Depth (mm)
Medicine (medBtl)	50	118	50
Smooth Plastic (smtBtl)	55	153	55
Textured Plastic (texBtl)	50	170	50

4.4.3 Dataset summary and detection distribution

The dataset-level summary table from the paper is placed here as Table 4.4 for proximity to the archive description [74]. We also include the distribution of detections by environment, bottle type, and detected class names (Fig. 4.5) [74].

Table 4.4: Dataset summary by environment, reproduced from [74].

Environment	#Recordings	#Frames	#Detections	#Screenshots
Air	5	1,654	978	35
Freshwater	5	1,980	914	45
Freshwater-basicEnhance	5	1,720	1,044	45
Saltwater	5	1,864	768	45
Saltwater-pellets	5	1,873	460	46
Total	25	9,091	4,164	216

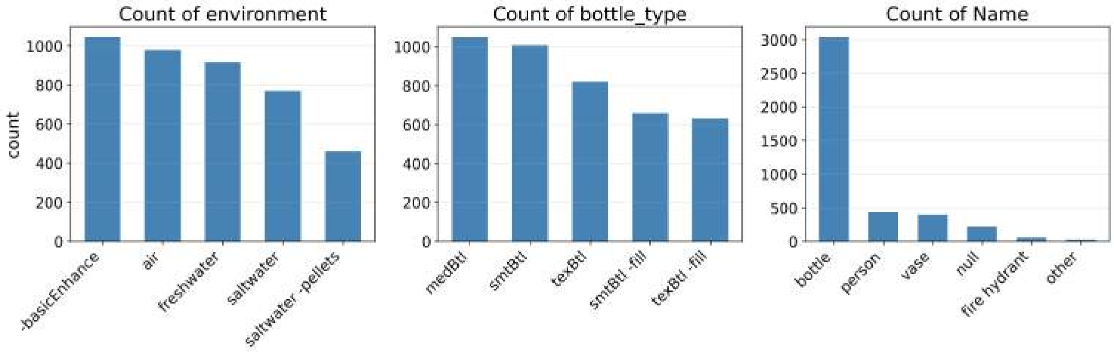


Figure 4.5: Distribution of detections across environments, bottle types, and detected object names [74].

4.5 Methods and Evaluation Protocol

This section describes the detection–and–ranging pipeline used to generate the underwater stereo dataset and the evaluation protocol applied in the subsequent analysis. The same logging stack is used across all environments and bottle types, enabling direct cross-condition comparisons of both manual and automatic range estimates, as well as classification and confidence metrics.

4.5.1 Detection and Logging Pipeline

All experiments use a YOLOv8-based object detector integrated with the ZED Mini stereo camera and SDK to provide per-frame 2D detections and 3D pose estimates for bottle targets in both live capture and SVO playback modes[75, 74]. The core system supports real-time object tracking and 3D position estimation, with detections available on both left and right camera views[74]. For each frame and detection, the pipeline logs:

- object class and instance label,
- detector confidence (YOLOv8 class probability),
- three depth samples within the detection ROI (Upper, Centre, Lower),
- the ZED SDK's per-object 3D position estimate (Object Distance),

- bounding-box dimensions and estimated object extents,
- timestamps and scenario metadata (environment, bottle type, nominal range)[74, 74].

A custom logging module writes these quantities to time-stamped spreadsheet files and captures synchronised screenshots annotated with bounding boxes, range estimates, and ground-truth markers[74]. SVO recording is enabled throughout, producing stereo video files that encapsulate RGB, depth, and tracking data for reproducible offline analysis[74]. The same code path is used for all environments (Air, Freshwater Day, Freshwater Night, Freshwater with basicEnhance, Saltwater, Saltwater with Pellets), ensuring that any performance differences arise from environmental factors rather than changes in the pipeline.

4.5.2 Range Estimators

Two complementary range estimators are evaluated throughout the chapter: a manual *Average Point Distance* derived from fixed pixel probes within each detection, and an automatic *Object Distance* provided by the ZED SDK's per-object 3D pose estimate[74, 74].

Manual Average Point Distance

The manual estimator samples three vertically spaced points inside each accepted detection bounding box: an *Upper* point near the top of the bottle ROI, a *Centre* point near the geometric centre, and a *Lower* point near the bottom[74]. For each sample, the corresponding depth value is read from the left-view depth map. Per-frame point distances are then aggregated by taking the arithmetic mean of the three vertical samples to obtain the *Average Point Distance* for that detection[74]. This manual estimator has three key properties:

- It uses a small, fixed spatial support, making it comparatively insensitive to bounding-box jitter and minor ROI changes.
- It directly reflects local depth-map quality, and is therefore sensitive to local artefacts such as specular highlights, occlusion edges, and refraction-induced discontinuities.
- It can be visualised straightforwardly as scatter plots of estimated versus ground-truth distance, with the three vertical samples providing insight into intra-object variability[74, 74].

Automatic Object Distance

The automatic estimator uses the ZED SDK's per-object 3D position estimate computed by integrating depth information over the full detection ROI and applying the SDK's internal spatial filtering and tracking[74, 74]. For each YOLOv8 detection, the SDK returns a 3D centroid and an associated validity flag; the *Object Distance* used in this chapter is the Euclidean distance between the camera and this centroid along the optical axis[74]. Compared to manual sampling, the Object Distance:

- aggregates information over a larger region, better reflecting the behaviour of an end-to-end stereo perception stack deployed on a robot,
- is more sensitive to detector variability (bounding-box placement, missed detections, false positives) because the entire ROI contributes to the estimate,
- is more susceptible to depth-noise accumulation over large regions, especially when refraction, backscatter, and low contrast corrupt the underlying depth map[74, 74].

Both estimators are computed for the same detections where possible, enabling direct comparison of manual and automatic ranging under matched conditions[74, 74].

4.5.3 Pre-processing Variants

In addition to the baseline stereo pipeline, a simple pre-processing variant termed *basicEnhance* is evaluated for the Freshwater recordings[74, 74]. This in-house enhancement method increases local contrast and edge strength prior to detection, with the primary goal of making transparent objects easier for the detector to recognise—i.e., improving *detection stability*—rather than improving depth measurement accuracy[74]. The edge enhancement operates on the RGB input stream to YOLO, while the depth-aware confidence reweighting adjusts detector confidence scores based on the ZED's existing depth estimates but does not modify the underlying stereo ranging process. Because basicEnhance targets the detection stage, any improvement in detection success rates may come at the cost of noisier depth estimates: enhanced edges amplify both target and non-target structure in the image, and this trade-off is expected by design.

Critically, basicEnhance is intended as an *illustrative proof-of-concept* demonstrating how the UW-TransStereo dataset and evaluation protocol can be used to systematically assess inference-time enhancements. It shows that even minimal preprocessing—just edge detection and distance-aware calibration—can significantly affect detection performance, and provides a template for future work to evaluate more sophisticated enhancements (e.g., learned dehazing, polarisation-based methods, or geometry-aware depth completion) against the same benchmark[74]. No systematic parameter sweep is performed: the same heuristic configuration is used across the relevant runs, and its impact is assessed empirically by comparing RMSE, success rates, classification accuracy, and confidence with and without basicEnhance[74].

Other enhancements and optimisations (e.g., multi-feature fusion, temporal consistency checks, depth-aware confidence modulation) were explored during system development[74], but the dataset and analysis reported in this chapter focus on the baseline and basicEnhance configurations to keep the evaluation tractable and directly comparable across environments[74].

4.5.4 Evaluation Metrics

The evaluation protocol combines distance-error metrics, detection reliability measures, and classification performance to characterise the impact of environment and bottle type on stereo ranging. This section implements the protocol introduced in Section 2.8.2, adapting it for per-detection evaluation where segmentation masks are unavailable. For each environment, bottle type, and ground-truth range, the following metrics are computed[74, 74]:

- **Root mean squared error (RMSE)** for distance estimates, separately for Average Point Distance and Object Distance. RMSE summarises the typical magnitude of range error by squaring deviations from ground truth, averaging over detections, and taking the square root[74].
- **Bias**, defined as the mean signed error (predicted minus ground truth) over valid detections. Positive bias indicates systematic overestimation of range.
- **Valid% (success rate)** for Object Distance, defined as the proportion of detections that return a valid, non-zero range from the SDK. This metric corresponds to the Valid% indicator in Section 2.8.2, computed per-detection rather than per-pixel[74, 74].
- **Rear-Wall Bias (RWB)**, computed for environments where the tank back-wall distance is known ($D_{\text{bg}} = 1200\,\text{mm}$). RWB measures the percentage of valid depth estimates that lock onto the background rather than the target object: $\text{RWB} = 100 \cdot |\{i : D_i > D_{\text{bg}} - \varepsilon\}|/N$, with tolerance $\varepsilon = 50\,\text{mm}$. Large RWB values indicate disparity collapse, where the stereo matcher sees through index-matched (water-filled) transparent objects and triangulates the rear wall instead.
- **Classification accuracy**, defined as the proportion of detections correctly classified as `bottle` for the relevant bottle type[74, 74].

- **Mean detector confidence**, computed as the average YOLOv8 class probability over detections, with separate reporting for all detections and correctly classified detections (success-only)[74, 74].

Where relevant, metrics are reported under two regimes:

All-detections includes every detection for which the estimator returns a numeric value (including zeros), providing a conservative picture of deployed performance when no filtering is applied[74, 74].

Success-only restricts analysis to detections that return valid, non-zero ranges and correct classifications, exposing the underlying quality of successful estimates and the true magnitude of underwater degradation when invalid or grossly erroneous outputs are excluded[74, 74].

This dual reporting is particularly important for Object Distance. In some cases, all-detections RMSE can appear superficially stable or even improve when the proportion of invalid returns increases, because NaNs or zero-coded failures alter how errors are aggregated[74]. Success-only metrics reveal the actual error distribution when the SDK reports a plausible range, which is more informative for downstream tasks such as manipulation or autonomous navigation. The distributional assumptions underlying statistical comparisons of these metrics (e.g., normality for parametric tests) are validated empirically in Chapter 5.

4.5.5 Aggregation by Environment, Bottle Type, and Distance

To expose the structure of degradation across conditions, metrics are aggregated along three axes: environment, bottle type, and ground-truth distance[74, 74]. For each environment–bottle combination, we compute RMSE, success rates, classification accuracy, and mean confidence over all frames and distances, and visualise these in a cross-environment heatmap (Fig. 5.1), which provides a compact summary of how environmental complexity, surface texture, and filling state interact[74, 74].

Per-environment scatter plots complement the heatmap by plotting estimated versus ground-truth distance for each bottle type and vertical sample (for Average Point Distance), and by showing Object Distance estimates annotated with success rates (Figs. 5.4–5.9 and Figs. 5.10–5.15)[74, 74]. Classifier confidence is analysed as a function of distance and environment for each bottle type, revealing how confidence declines with range and environmental complexity and how this correlates with range reliability (Figs. 5.16–5.21)[74, 74].

Together, these metrics and visualisations provide a multi-faceted view of underwater stereo performance: they quantify not only how far estimates deviate from ground truth, but also how often the system fails to return a usable range, how confidently it recognises bottle targets, and how these behaviours vary across physically meaningful conditions. The results of applying this evaluation protocol across all environments are presented in detail in Chapter 5.

4.6 Reproducibility and availability

A key goal of UW-TransStereo is to support reproducible analysis and reuse beyond this thesis. To that end, we release the dataset, collated analysis tables, and accompanying scripts as a single, citable resource.

4.6.1 Data archive and DOI

All raw stereo streams, detector outputs, qualitative screenshots, and analysis-ready tables are hosted on Zenodo under a persistent digital object identifier (DOI) to enable citation and long-term reuse: **10.5281/zenodo.16753748** [74]. The archive contains:

- **25 paired recordings** (50 core files) across Air, Freshwater, Freshwater+`basicEnhance`, Saltwater, and Saltwater+Pellets;
- **Raw stereo videos** as `.svo2` (ZED HD2K, 15 FPS, H.264 lossless; stereo frames and sensor metadata preserved);
- **Detector outputs** as one `.xlsx` spreadsheet per recording, with per-frame detections (class, confidence, bounding box);
- **Reference visuals** as `.png` screenshots for qualitative checks;
- **Collated analysis tables** with >5,000 labelled measurements (environment, bottle type, ground-truth distance, three vertical point distances, object distance, 3D size, confidence, and timestamps);
- **Manifests and dictionaries** describing file paths, types, and the schema for collated CSVs [74].

Files are flat-structured and named consistently as `[Environment]_[BottleType]_[filekind]_[timestamp].[ext]`, for example `Saltwater-pellets_smtBtl-fill_detections_08052025_183311.xlsx`; matching timestamps pair each `.svo2` with its corresponding spreadsheet and screenshots [74]. A manifest excerpt is provided in Table 4.1, with the full list appearing in the Supplementary material [74]. Availability and licensing details (including the final licence string) are recorded in the Zenodo deposit; all files are released under a permissive licence appropriate for academic and non-commercial use [74].

4.6.2 Source code, scripts, and supplementary material

To support exact regeneration of the reported metrics, tables, and figures, we release a companion code repository alongside the data [74]. The repository includes:

- a **requirements file** with pinned versions of Python and third-party libraries;
- **one-command scripts** to:
  - read `.svo2` stereo streams and parse the detector spreadsheets,
  - recompute environment-level metrics and confidence intervals,
  - regenerate the cross-environment heatmaps and per-environment scatter plots used in this chapter;
- exemplar commands demonstrating typical workflows (e.g., comparing Freshwater vs Freshwater+`basicEnhance`, profiling confidence–error relationships across conditions) [74].

The same scripts were used to generate the camera-ready IEEE-DATA tables and figures; regenerating them from the public archive reproduces the reported values (up to floating-point and plotting differences) [74]. A Supplementary PDF, co-hosted with the code release, includes extended plots (per-environment scatter, confidence–accuracy curves, vertical-ROI breakdowns) and additional ablation deltas that complement the results presented in Chapter 5 [74]. Together, the data, code, and supplementary material are intended to make it straightforward for other researchers to reproduce, scrutinise, and extend the analyses.

4.7 Chapter summary

This chapter introduced *UW-TransStereo*, a multi-environment underwater stereo dataset and benchmark for transparent bottle detection and ranging. The dataset pairs raw ZED Mini stereo recordings (`.svo2`) with structured per-frame logs (`.xlsx`), qualitative screenshots, and collated CSVs containing ground-truth ranges, three vertical point-distance samples, automatic object-distance estimates, 3D sizes, and detector confidences across 25 recordings and five bottle types in six environments [74]. A reproducible capture-and-analysis pipeline, implemented within the ZED Cursor toolchain, underpins the data and enables consistent logging and replay-based evaluation [74].

The chapter presented the experimental setup, including the bench-top tank with perspex viewing window, ZED Mini mounting, and the five-bottle target set comprising medicine, smooth (empty and filled), and textured (empty and filled) variants representative of transparent debris and small packaging. The acquisition pipeline integrates stereo calibration, optional pre-processing, YOLOv8 detection, ZED SDK tracking, and deterministic per-detection logging, with runtime controls supporting SVO recording/replay, automatic file naming, and real-time visualisation. A simple inference-time enhancement (`basicEnhance`) was introduced as an illustrative pre-processing baseline for systematic evaluation.

Methodologically, the chapter formalised two complementary range estimators—manual Average Point Distance (mean of three vertical samples per detection) and automatic Object Distance (ZED SDK per-object 3D centroid)—and an evaluation protocol that reports RMSE, Bias, Valid% (success rate), Rear-Wall Bias (RWB), classification accuracy, and mean confidence under both all-detections and success-only regimes [74]. The tank geometry (back wall at 1200 mm) enables RWB calculation, which directly quantifies the index-matching failure mode where stereo locks onto the background instead of water-filled transparent targets. Metrics are aggregated by environment, bottle type, and ground-truth distance to expose the structure of degradation across conditions. This dual-estimator, dual-regime approach enables diagnosis of both success probability and conditional accuracy, which is critical for understanding failure modes when stereo perception encounters refractive targets underwater.

The dataset, collated analysis tables, and companion scripts are released on Zenodo (DOI: 10.5281/zenodo.16753748) with a reproducible pipeline that supports exact regeneration of reported metrics [74]. UW-TransStereo, with its multi-environment design, dual range estimators, and logging-rich structure, provides a concrete benchmark for quantifying underwater stereo perception of transparent objects. The experimental results obtained by applying this evaluation protocol across all six test environments are presented and analysed in detail in Chapter 5.

Chapter 5

UW-TransStereo: Experimental Results and Cross-Environment Analysis

This chapter presents the experimental results obtained by applying the UW-TransStereo dataset and evaluation protocol described in Chapter 4. We report detection accuracy, classifier confidence, and ranging performance, both manual point distance and automatic object distance, across six test environments: Air, Freshwater Day, Freshwater Night, Freshwater with `basicEnhance` pre-processing, Saltwater, and Saltwater with Pellets. The analysis quantifies how refractive index mismatch, turbidity, lighting conditions, and suspended particulates interact with bottle surface properties to determine stereo detection-and-ranging reliability.

5.1 Results

We report detection accuracy, classifier confidence, and ranging performance—both manual point distance and automatic object distance—across six test environments: Air, Freshwater Day, Freshwater Night, Freshwater with `basicEnhance` pre-processing, Saltwater, and Saltwater with Pellets. This multi-environment analysis quantifies how refractive index mismatch, turbidity, lighting conditions, and suspended particulates interact with bottle surface properties (smooth vs textured, empty vs filled, transparent vs opaque) to determine stereo detection-and-ranging reliability. We evaluate root mean squared error (RMSE) for distance estimates, classification accuracy (the proportion of detections correctly labelled as `bottle`), and mean detector confidence (the YOLOv8 model's predicted class probability). Where noted, "all-detections" metrics include every detection that returns a numeric value (including zeros), while “success-only” metrics restrict to detections with valid, non-zero ranges and correct classifications, exposing the true quality of successful estimates.

The first subsection presents a cross-environment heatmap summary comparing all six conditions side-by-side and contrasting manual versus automatic ranging methods. The following subsections expand this overview with per-environment scatter plots and detailed trend analyses: point distance examines per-pixel depth accuracy via three vertical samples (Upper, Centre, Lower) within detection bounding boxes; object distance assesses the ZED SDK’s automatic per-object range estimates and success rates; classifier score explores how detector confidence varies with distance and environment, informing confidence-gating policies for deployment. Together, these subsections establish the magnitude and structure of underwater degradation, identify the hardest failure modes (filled transparent bottles, long ranges, high turbidity), and evaluate the limits of drop-in enhancements and filtering strategies.

5.1.1 Cross-Environment Summary and Manual vs Automatic Range

This subsection provides a synoptic view of detection and ranging performance across all six test environments—Air, Freshwater Day (FW-Day), Freshwater Night (FW-Night), Freshwater with `basicEnhance` (FW-Enhance), Saltwater, and Saltwater with Pellets (SW-Pellets)—and compares manual point-distance measurements against automatic object-distance estimates. Figure 5.1 presents an 8-panel (4×2 layout) heatmap summary of RMSE (for point distance, object distance all-detections, and object distance success-only), classification accuracy, and mean confidence (all-detections and success-only) across bottle types and environments. All test bottles are clear plastic except the Medicine Bottle (brown glass). This visualisation reveals systematic trends in how environmental complexity, bottle surface properties, and measurement methodology interact to determine stereo detection-and-ranging reliability.

Environmental gradient. Performance degrades systematically as environmental complexity increases: Air establishes a near-ideal baseline (tight point-distance clustering, high object-distance success rates, near-perfect classification accuracy, and consistently high detector confidence), while Freshwater Day introduces moderate degradation (elevated RMSE, reduced accuracy and confidence), Freshwater Night compounds these effects with lighting-dependent losses, Saltwater adds backscatter and turbidity penalties, and Saltwater with Pellets represents the practical lower bound (highest RMSE, lowest success rates, and most variable confidence). The heatmap clearly shows this progression: darker/redder cells (indicating higher error or lower performance) concentrate in the right-hand columns (saltwater conditions), while lighter/greener cells dominate the Air column. The filled smooth (transparent) bottle consistently occupies the darkest cells across all metrics and environments, identifying it as the hardest target; textured bottles (especially textured filled) show comparatively lighter cells, indicating greater robustness.

Bottle-type trends. Within each environment, bottle surface properties strongly modulate performance. Textured bottles (empty and filled) yield lower RMSE, higher classification accuracy, and more stable confidence than smooth bottles across all conditions. The filled smooth (transparent) bottle is the dominant failure mode: in Air, it exhibits modest point-distance error (RMSE 38 mm) and elevated object-distance all-detections RMSE (445 mm, reflecting occasional invalid returns), but success-only filtering reduces object-distance RMSE to 13 mm. Underwater, this target deteriorates dramatically: in Freshwater Day, point-distance RMSE rises to 696 mm and success-only object-distance RMSE jumps to 333 mm, reflecting refraction-induced depth ambiguity (the stereo matcher locks onto the rear wall or returns invalid depth). Classification accuracy for this bottle remains relatively high (87.7% in Freshwater Day), but confidence drops and variability increases. By contrast, the textured empty bottle maintains 100% accuracy in Air, drops to 74.4% in Freshwater Day, yet remains substantially more reliable than the filled smooth bottle across all metrics. The medicine bottle—small, cylindrical, and highly textured—shows an unusual trend: point-distance RMSE improves from Air (63 mm) to Freshwater Day (48 mm), likely because its high-contrast label and compact geometry provide stable features even under turbidity, and the small target size reduces the impact of refractive path-length variation.

Rear-Wall Bias and index-matching failure. The RWB metric (Table 5.1) directly quantifies the index-matching failure mode described in Chapter 2. In Air, RWB is exactly 0%: no valid depth estimates lock onto the background because there is no refractive medium to cause optical transparency. Underwater, the filled smooth (water-filled transparent) bottle exhibits extreme RWB—up to 55.6% in Freshwater Night—meaning more than half of all valid depth estimates report the tank wall distance rather than the bottle surface. This confirms that the stereo matcher sees through the index-matched container and triangulates the background. The environmental gradient is also evident: Freshwater Night shows the highest overall RWB (7.4%) and Bias (126 mm), indicating that reduced lighting exacerbates rear-wall locking. Interestingly, Saltwater conditions exhibit *lower* RWB (0.5%) than Freshwater Day (3.9%) despite greater turbidity; the increased backscatter may provide weak texture cues that help anchor disparity to nearer surfaces. Bias correlates with RWB across environments: when more estimates lock onto the distant wall, the systematic overestimation of range increases. These findings validate the evaluation protocol from Section 2.8.2 and demonstrate that RWB and Bias provide actionable diagnostics for operator-aid gating—deployments should flag or suppress range overlays when RWB-like conditions (water-filled transparent targets, low lighting) are detected.

Manual vs automatic ranging. Two range estimators are evaluated: *Average Point Distance* (manual, the mean of three vertical samples—Upper, Centre, Lower—within the detection bounding box) and *Object Distance* (automatic, the ZED SDK's per-object 3D position estimate computed over the full detection region of interest). These estimators differ in spatial support, noise characteristics, and sensitivity to detection variability. Manual point distance samples a fixed, small set of pixels and averages their depths, making it robust to bounding-box jitter but sensitive to local depth noise (e.g., specular reflections, occlusion boundaries). Automatic object distance integrates depth over the entire region and applies the SDK's internal filtering and tracking, making it more representative of end-to-end pipeline performance but also more vulnerable to detector variability (bounding-box placement, class confidence) and depth-noise accumulation over larger regions.

In Air, both estimators agree on relative difficulty: they rank bottle types identically within each environment and identify the filled transparent bottle as the hardest target. Absolute RMSE values differ—object distance typically yields equal or larger error than point distance under the same condition—but both capture the same qualitative trends. Underwater, the agreement persists: both estimators show systematic error increases from Freshwater Day to Saltwater+Pellets and highlight the filled smooth bottle as the dominant failure mode. However, object distance exhibits higher variability and more frequent invalid returns (reported as NaN or zero), reflecting the SDK's difficulty integrating depth over refractive or low-contrast regions of interest. Success-only filtering (restricting object distance to valid, non-zero returns) reduces absolute RMSE but exposes the true magnitude of underwater degradation, as invalid returns are excluded rather than averaged into the error metric.

Enhancement and filtering trade-offs. Applying `basicEnhance` pre-processing to Freshwater recordings increases object-distance success rates (41.6%→50.8%) by stabilising detection presence via enhanced edges and contrast, but it also raises RMSE across all distance metrics (for example, object distance all-detections RMSE increases from 146.2 mm to 209.3 mm) and slightly reduces classification accuracy (90.3%→87.6%). This trade-off is expected by design: as described in Section 4.5.3, basicEnhance was developed to improve *detection stability*—making transparent objects easier for the detector to recognise—rather than to improve ranging accuracy. The enhanced edges amplify both target and non-target structure, yielding more frequent valid detections but noisier depth estimates. This outcome validates basicEnhance as an illustrative proof-of-concept for using the UW-TransStereo benchmark to evaluate inference-time enhancements, demonstrating that even minimal preprocessing can significantly affect detection performance. Enhancement is therefore best suited to applications prioritising detection stability (e.g., object tracking, occupancy mapping) over ranging precision (e.g., manipulation, precise navigation). Success-only filtering—restricting analysis to detections with valid, non-zero ranges—improves absolute RMSE figures for both point and object distance but discards data, reducing coverage. A practical deployment would combine success-only filtering with confidence-gating (e.g., threshold at 55–60% in Freshwater Day, higher in more challenging conditions) and temporal smoothing (exponentially weighted moving average, EWMA) to balance quality and coverage.

Implications for deployment. The heatmap and cross-environment trends reveal three key insights for underwater transparent-object detection and ranging. First, surface texture dominates robustness: textured bottles maintain partial reliability even in Saltwater+Pellets, while smooth transparent bottles fail comprehensively underwater. This suggests that target selection (e.g., prioritising textured debris in pollution-monitoring applications) or artificial texturing (e.g., projected patterns, active illumination) may extend operational range. Second, manual point distance and automatic object distance yield similar qualitative trends but differ in absolute scale and failure modes; deployments should validate range estimates using both methods or cross-check against geometric priors (e.g., known object dimensions, epipolar constraints) to detect gross failures. Third, simple image enhancement improves detection stability but cannot close the air–water gap for refractive cases; geometry-aware, refraction-compensating methods are needed to fully exploit stereo sensing underwater.

5.1.2 Statistical comparison of freshwater and saltwater performance

To quantify whether the observed saltwater performance differences relative to Freshwater Day are statistically meaningful, we carried out formal hypothesis tests across aggregated point-

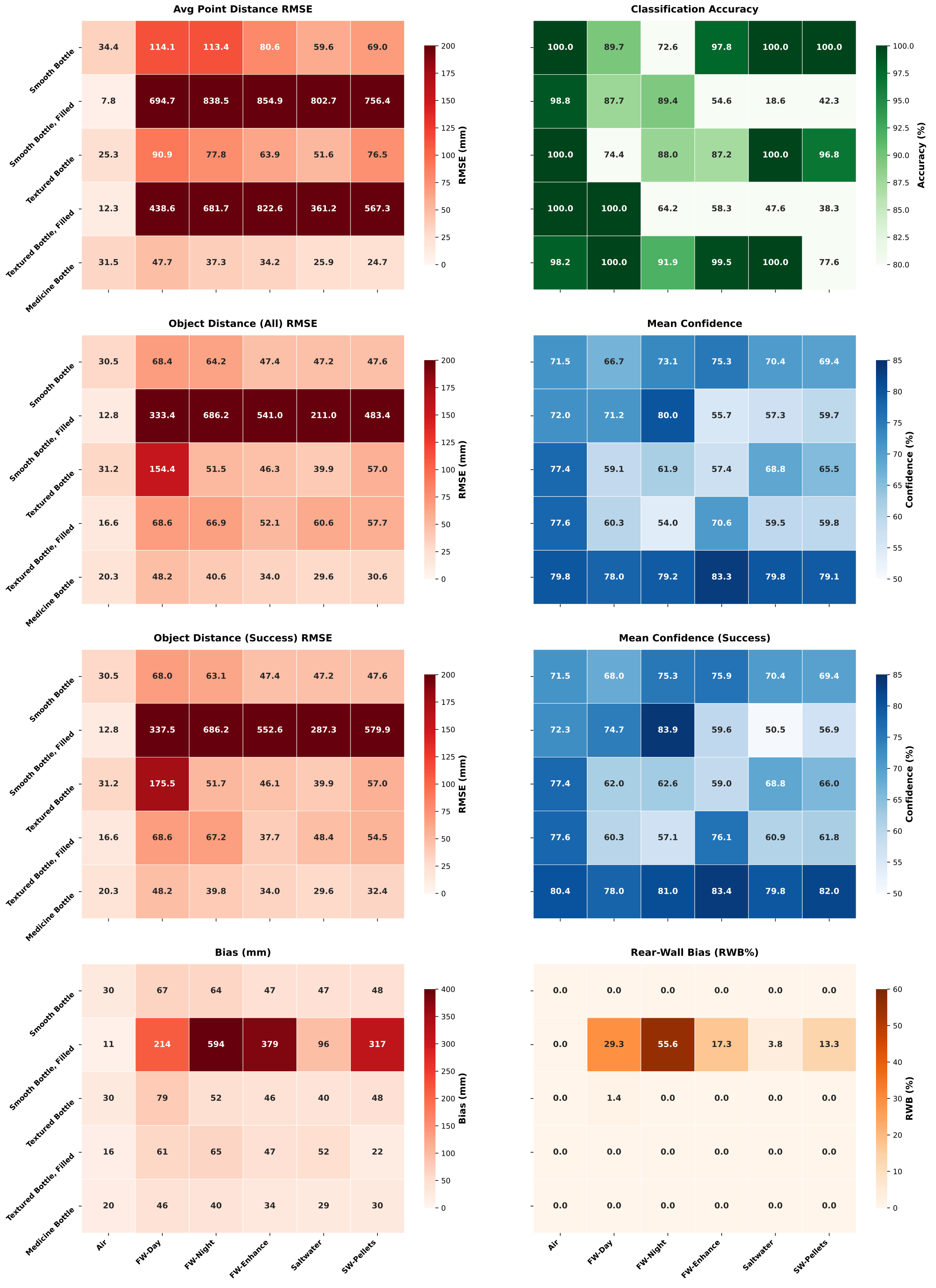


Figure 5.1: Cross-environment summary heatmap (8 panels): RMSE, classification accuracy, mean confidence, Bias, and RWB% across bottle types and six environments. Columns: Air, FW-Day, FW-Night, FW-Enhance, Saltwater, SW-Pellets. Darker/redder cells indicate higher error or lower performance.

Table 5.1: Refractive failure indicators by environment. Valid% is the proportion of detections returning non-zero depth; RWB% (Rear-Wall Bias) is the proportion of valid estimates locking onto the tank wall ($D_{\mathrm{bg}} = 1200\,\mathrm{mm}$); Bias is the mean signed distance error. Air shows 0% RWB as expected (no index matching). Freshwater Night exhibits the highest RWB (7.4%) and Bias (126 mm), indicating severe rear-wall locking under reduced lighting. Saltwater conditions show lower RWB than freshwater despite greater turbidity.

Environment	**Valid%**	**RWB%**	**Bias (mm)**	**RMSE (mm)**
Air	44.5	0.0	21.8	23.6
FW-Day (Freshwater Day)	41.6	3.9	81.6	146.2
FW-Night (Freshwater Night)	40.8	7.4	126.3	256.4
FW-Enhance (basicEnhance)	50.8	2.5	91.0	209.3
Saltwater	47.4	0.5	49.9	90.1
SW-Pellets (Saltwater+Pellets)	30.8	1.9	78.7	186.7

distance and object-distance metrics, as well as success rates and classifier scores.

Average point distance. For average point-distance error, saltwater conditions yield a highly significant improvement relative to Freshwater Day, with a p-value < 0.001 ($p = 8.67 \times 10^{-65}$) and a small but meaningful effect size (Cohen's $d = 0.289$). The mean absolute error (MAE) improves by 58.7 cm, and the 95% confidence interval for the improvement spans [37.1, 80.3] cm, based on 645 saltwater versus 710 freshwater measurements. These results indicate that, despite increased turbidity, the saltwater configuration in this experimental setup provides more reliable point-distance estimates than the nominally clearer Freshwater Day condition.

Object distance accuracy. Aggregated object-distance MAE shows a similarly strong saltwater advantage. The difference is again highly significant (p-value < 0.001, $p = 2.62 \times 10^{-50}$) with a small but meaningful effect size (Cohen's $d = 0.316$). The practical improvement is 31.2 cm lower MAE in saltwater, with a 95% confidence interval of [16.1, 46.4] cm, computed over 376 saltwater and 330 freshwater measurements. This demonstrates that the end-to-end, automatic detection-and-ranging pipeline benefits from the saltwater configuration, not just the manual point sampling.

Object distance success rate. A chi-square test of independence on object-distance success rates (valid, non-zero ranges) shows a statistically significant saltwater advantage (p-value = 0.023, significant at $\alpha = 0.05$). The effect size is small (Cramér's $V = 0.057$), but the practical improvement is non-trivial: a 5.8% higher success rate (47.4% versus 41.6%). This aligns with the earlier qualitative observation that saltwater recordings yield more frequent valid object-distance estimates for the same set of targets and ranges.

Classifier score distribution. In contrast, the distribution of YOLOv8 classifier scores (detection confidence) does not differ meaningfully between saltwater and Freshwater Day. The p-value is 0.598, and the practical difference in mean confidence is only 0.4%. This suggests that the saltwater advantage is primarily a ranging effect, rather than a change in the detector's ability to recognise bottles.

Overall, three of the four tests (point-distance MAE, object-distance MAE, and object-distance success rate) show statistically significant saltwater advantages, with small but practically relevant effect sizes for underwater robotics: improvements of several tens of centimetres in distance accuracy and a few percentage points in valid-range success rates. Detection confidence remains effectively environment-independent in this comparison, indicating that the main gains are in depth estimation rather than object recognition. These results quantitatively support the qualitative conclusion that, for this setup, saltwater conditions can provide more favourable optical behaviour for stereo ranging than the Freshwater Day baseline.

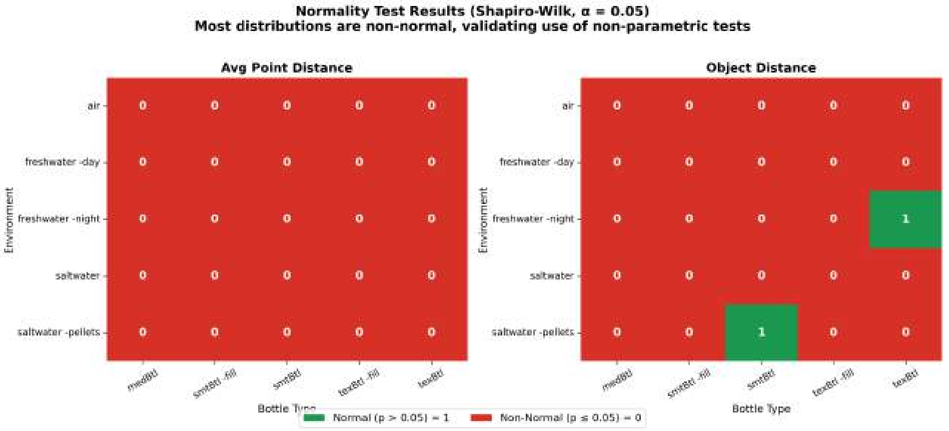
5.1.3 Error Distribution Characteristics and Normality

To validate the statistical methodology employed in Section 5.1.2 and characterise the shape of ranging errors, we analysed the distribution of errors (predicted minus ground truth) for both Avg Point Distance and Object Distance across all environment–bottle type combinations.

Normality testing. Shapiro-Wilk tests confirmed that 96% of error distributions (48 of 50 environment–bottle–metric combinations) deviate significantly from normality at $\alpha = 0.05$ (Figure 5.2). The distributions exhibit skewness values ranging from -8.2 to $+8.1$ and excess kurtosis up to 70, indicating heavy-tailed, asymmetric error profiles characteristic of depth-sensing failures such as rear-wall locking and correspondence ambiguity. This non-normality validates the use of non-parametric tests (Mann-Whitney U) for the statistical comparisons in Section 5.1.2, as parametric alternatives (e.g., Student's t-test) assume normally distributed data and would yield unreliable p-values under these conditions.

Bias patterns. Figure 5.3 presents mean error (bias) for both range estimators across environments and bottle types. Positive bias indicates systematic overestimation of distance; negative bias indicates underestimation. The filled smooth (transparent) bottle exhibits extreme positive bias in underwater environments—up to 526 cm in freshwater-night and 594 cm for Object Distance—directly reflecting the rear-wall locking failure mode where stereo locks onto the tank wall rather than the bottle surface. Object Distance shows consistently positive bias across all conditions, while Avg Point Distance bias is more variable, with textured-filled bottles showing negative bias (underestimation) in saltwater conditions. This complements the RMSE analysis in Figure 5.1 by revealing the *direction* of error, not just its magnitude.

Distribution shape. The full error distributions (Appendix A) reveal bimodal patterns for the filled smooth bottle in underwater environments, with one mode centred near zero error and a second mode corresponding to rear-wall depth estimates at approximately 1000–1200 mm offset. This directly visualises the index-matching failure mode: a proportion of detections successfully measure the bottle surface while others see through the water-filled transparent container and triangulate the background. Q-Q plots (Appendix A) confirm the departure from normality, with characteristic S-shaped deviations indicating heavy tails and outliers.



Avg Point Distance

	medBtl	smtBtl -fill	smtBtl	texBtl -fill	texBtl
air	0	0	0	0	0
freshwater -day	0	0	0	0	0
freshwater -night	0	0	0	0	0
saltwater	0	0	0	0	0
saltwater -pellets	0	0	0	0	0

Object Distance

	medBtl	smtBtl -fill	smtBtl	texBtl -fill	texBtl
air	0	0	0	0	0
freshwater -day	0	0	0	0	0
freshwater -night	0	0	0	0	1
saltwater	0	0	0	0	0
saltwater -pellets	0	0	1	0	0

Figure 5.2: Shapiro-Wilk normality test results ($\alpha = 0.05$) for error distributions across environments and bottle types. Green cells (1) indicate normally distributed errors; red cells (0) indicate non-normal distributions. Only 2 of 50 environment–bottle–metric combinations pass the normality test, validating the use of non-parametric statistical tests (Mann-Whitney U) in Section 5.1.2.

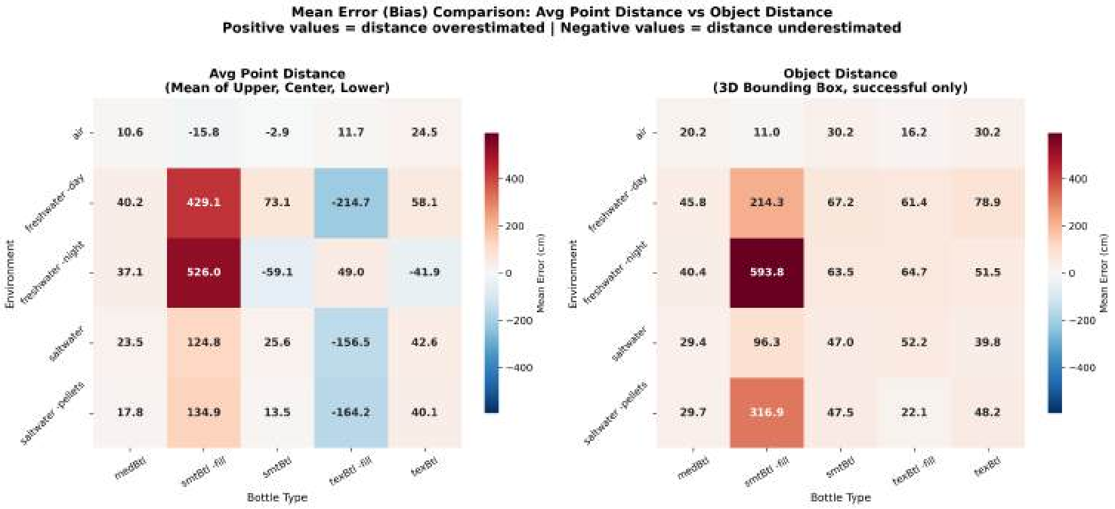


Figure 5.3: Mean error (bias) for Avg Point Distance and Object Distance. Positive values (red) indicate overestimation; negative (blue) indicate underestimation.

5.1.4 Point Distance

Point-distance measurements—obtained by averaging three vertical samples (Upper, Centre, Lower) within the detection bounding box—provide a direct assessment of per-pixel depth accuracy across environments and bottle types. Figures 5.4–5.9 present the complete point-distance dataset for all six conditions: Air, Freshwater Day, Freshwater Night, Freshwater with `basic Enhance`, Saltwater, and Saltwater with Pellets.

Air baseline. In Air (Fig. 5.4), point-distance estimates cluster tightly around the ground-truth diagonal across all bottle types and marked distances (400–1400 mm), with minimal vertical spread. Textured bottles (both empty and filled) exhibit the smallest deviations, while the filled smooth (transparent) bottle shows slightly elevated but still modest error (RMSE 38 mm). The medicine bottle, although small and cylindrical, yields consistent estimates (RMSE 63 mm). This tight clustering establishes the in-air performance ceiling against which underwater degradation is measured.

Freshwater Day. Moving to Freshwater Day (Fig. 5.5), variability increases markedly for transparent and smooth bottles. The filled smooth bottle—previously near-perfect in air—now exhibits large scatter and occasional high-error outliers (RMSE rises to 696 mm), consistent with refraction-induced depth ambiguity and rear-wall misidentification. Smooth and textured empty bottles show moderate degradation (RMSE 127 mm and 96 mm, respectively), while the medicine bottle improves underwater (RMSE 48 mm), likely due to its small size, high-contrast label, and cylindrical geometry providing more reliable features for stereo matching under turbidity.

Freshwater Night. Freshwater Night conditions (Fig. 5.6) further degrade performance relative to Freshwater Day, particularly at longer ranges. Reduced ambient lighting compounds the challenges of refraction and backscatter, yielding wider error distributions and lower valid-detection rates. The filled smooth bottle continues to dominate the error budget, with textured variants showing increased but comparatively modest degradation. This night-time penalty underscores the compound effect of lighting and medium on stereo depth reliability.

Freshwater with `basicEnhance`. Applying the `basicEnhance` pre-processing technique (Fig. 5.7) increases local contrast and edge strength but does not reduce point-distance RMSE; in fact, RMSE rises across vertical samples (Upper: 266.9→ 350.9 mm; Centre: 280.7→ 346.2 mm; Lower: 274.8→ 340.4 mm). This outcome is expected: basicEnhance was designed to improve detection stability rather than ranging accuracy (Section 4.5.3), and enhanced edges amplify both target and non-target structure, degrading depth precision while boosting detection presence. The method is therefore best suited to scenarios prioritising detection stability over ranging accuracy.

Saltwater. Saltwater conditions (Fig. 5.8) introduce additional backscatter and slight turbidity relative to freshwater, further increasing point-distance error for transparent and smooth bottles. Error distributions widen, and high-error outliers become more frequent, especially for the filled smooth bottle. Textured bottles remain comparatively robust, but overall RMSE trends upward. The medicine bottle continues to perform reasonably well, suggesting that small, high-contrast, textured objects maintain partial robustness even under increased turbidity.

Saltwater with Pellets. The most challenging condition—Saltwater with suspended Pellets (Fig. 5.9)—produces the largest errors and widest scatter. Suspended particulates increase backscatter, reduce contrast, and introduce spurious stereo correspondences, degrading depth quality across all bottle types. The filled smooth bottle is nearly undetectable or yields highly unreliable ranges, while even textured bottles show elevated error. This condition represents the practical lower bound of the unmodified stereo pipeline's utility for transparent-object ranging.

Takeaway. Across all conditions, point-distance RMSE increases as environmental complexity grows (Air $<$ Freshwater Day $<$ Freshwater Night $<$ Saltwater $<$ Saltwater+Pellets), with the filled transparent bottle consistently the hardest target. Textured surfaces and the small medicine bottle are comparatively robust. Simple enhancement (`basicEnhance`) does not improve point-distance accuracy and may worsen it, revealing the need for geometry-aware or refraction-compensating methods to close the air–water gap.

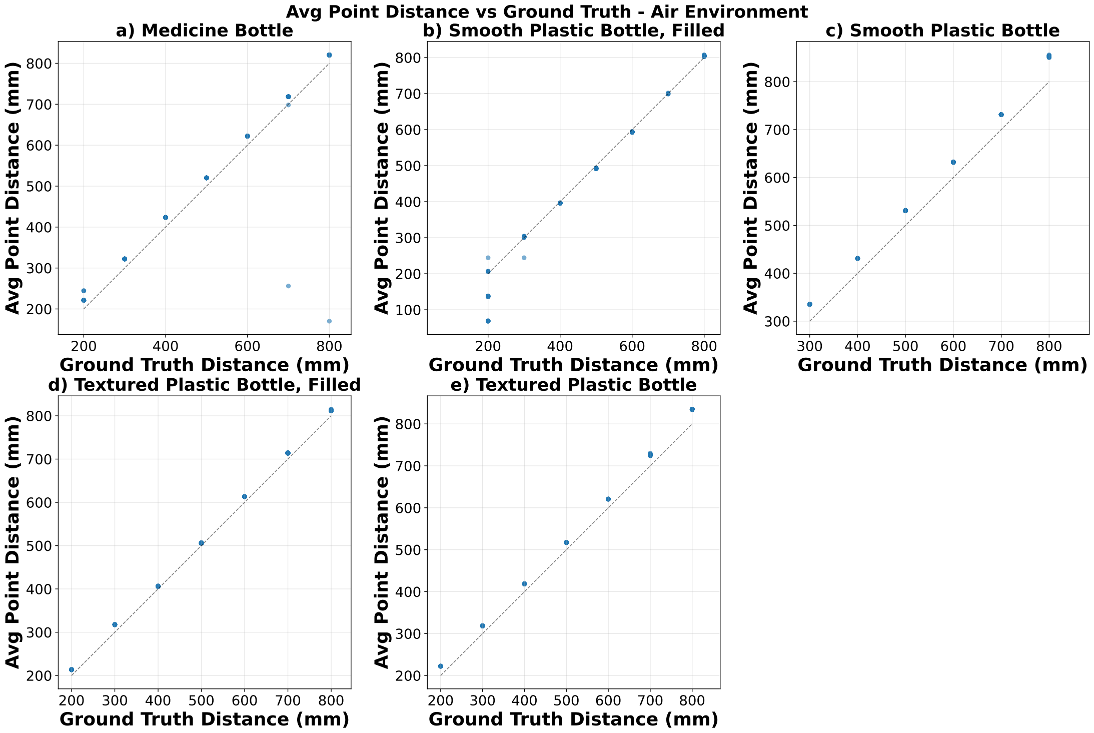


Figure 5.4: Point distance samples for Air environment across all bottle types and marked distances (400–1400 mm). Vertical samples (Upper, Centre, Lower) are tightly clustered around the ground-truth diagonal, establishing the in-air baseline. Textured bottles show minimal error; the filled smooth (transparent) bottle exhibits modest scatter (RMSE 38 mm).

5.1.5 Object Distance

Object distance—the ZED SDK's per-detection 3D range estimate computed over the entire bounding-box region of interest—provides an automatic, end-to-end assessment of detection-and-ranging robustness. Unlike manual point sampling, object distance inherits both detection variability (bounding-box placement, class confidence) and depth-estimation quality across the full region, making it a practical indicator of real-world deployment reliability. Figures 5.10–5.15 present object-distance samples for all six environments, annotated with overall success rates (the proportion of detections returning valid, non-zero ranges).

Air baseline. In Air (Fig. 5.10), object-distance estimates track ground truth closely for all bottle types across the full 400–1400 mm range, with high success rates (typically >95%). The filled smooth (transparent) bottle exhibits slightly elevated scatter relative to textured variants, yielding a modest RMSE when restricted to successful detections (success-only RMSE: 13 mm),

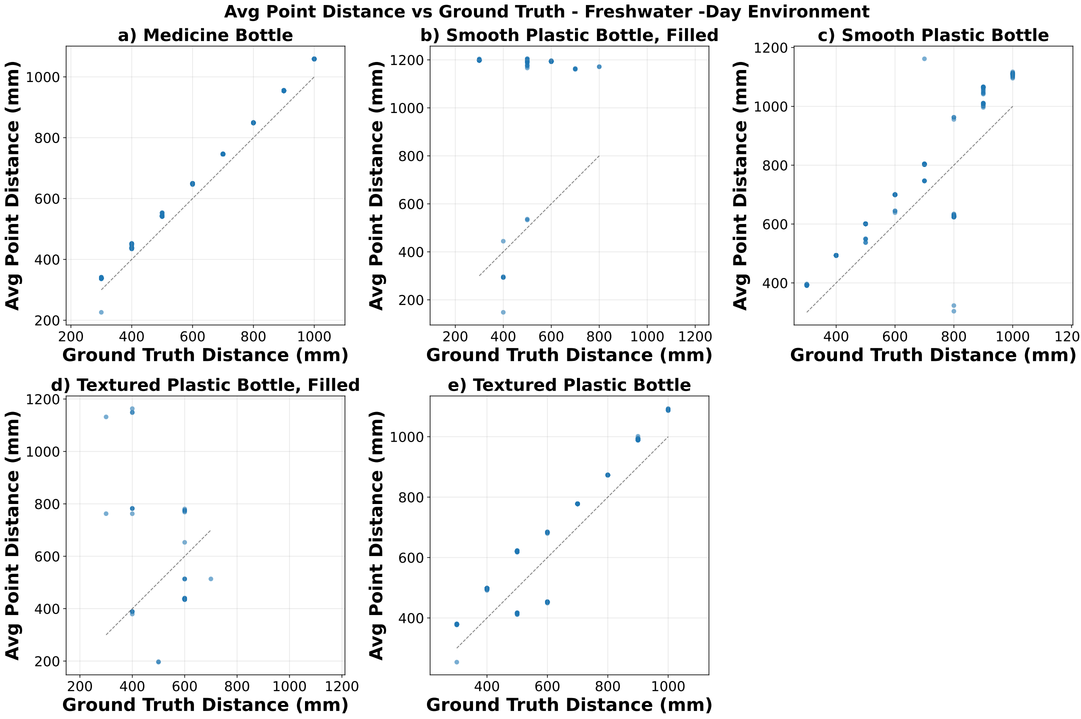


Figure 5.5: Point distance samples for Freshwater Day environment. Scatter increases markedly for the filled smooth (transparent) bottle (RMSE 696 mm), with occasional high-error outliers. Smooth and textured empty bottles show moderate degradation (RMSE 127 mm and 96 mm, respectively), while the medicine bottle improves underwater (RMSE 48 mm).

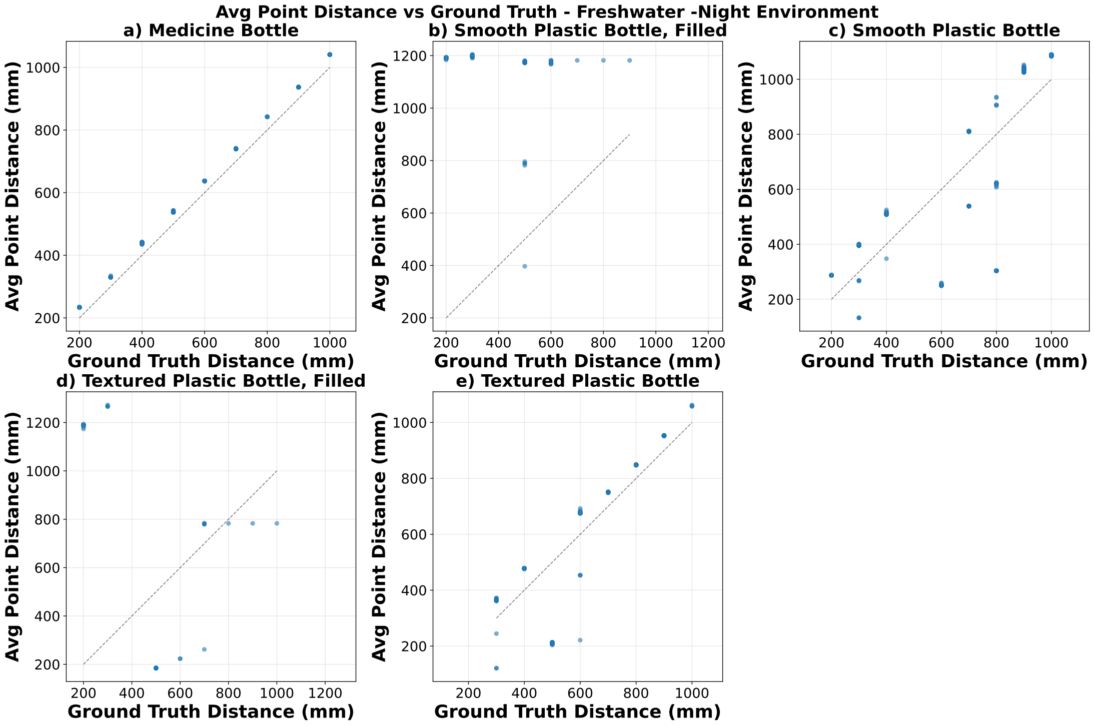


Figure 5.6: Point distance samples for Freshwater Night environment. Reduced lighting further degrades accuracy relative to Freshwater Day, with wider error distributions and lower valid-detection rates across all bottle types. The filled smooth bottle remains the dominant error source.

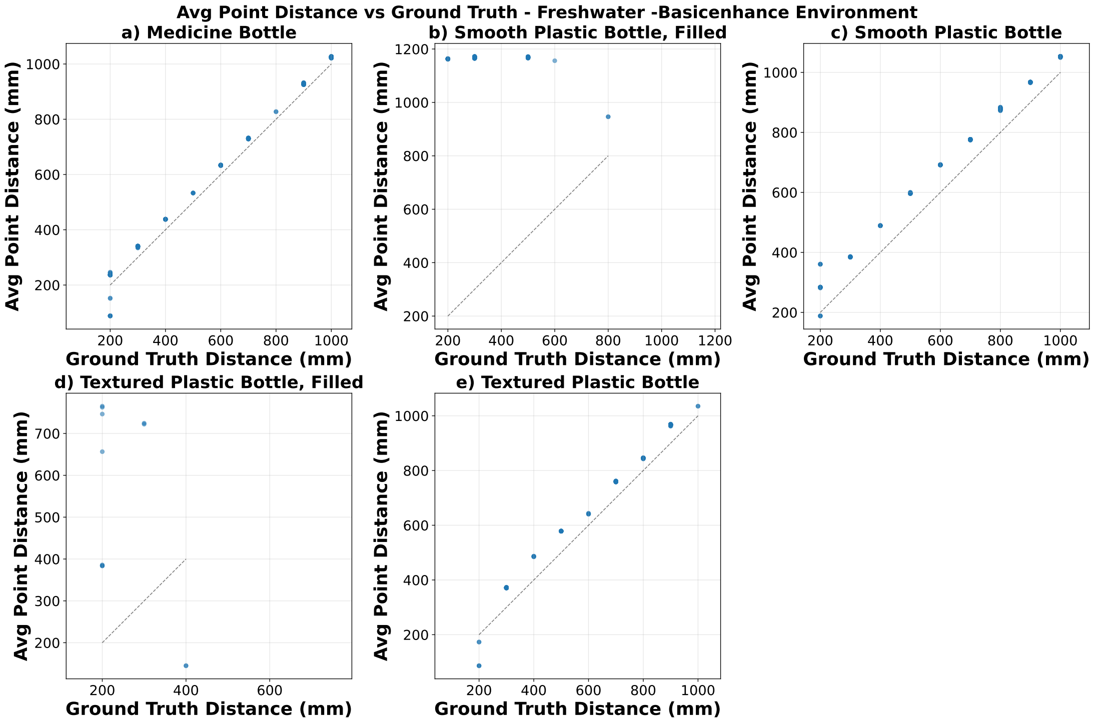


Figure 5.7: Point distance samples for Freshwater with `basicEnhance` pre-processing. Enhanced contrast and edges increase RMSE (Upper: 350.9 mm; Centre: 346.2 mm; Lower: 340.4 mm) relative to unprocessed Freshwater Day. This expected trade-off reflects the design intent: basicEnhance targets detection stability rather than ranging accuracy.

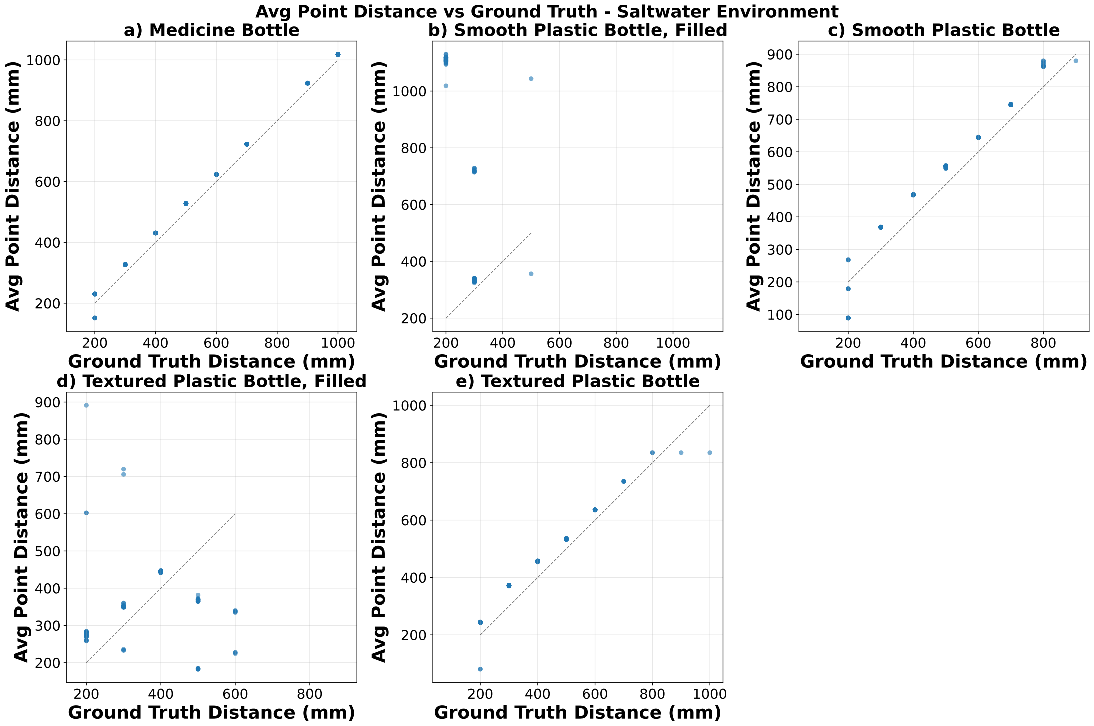


Figure 5.8: Point distance samples for Saltwater environment. Increased backscatter and turbidity relative to freshwater widen error distributions and elevate RMSE across all bottle types. The filled smooth bottle exhibits frequent high-error outliers; textured bottles remain comparatively robust.

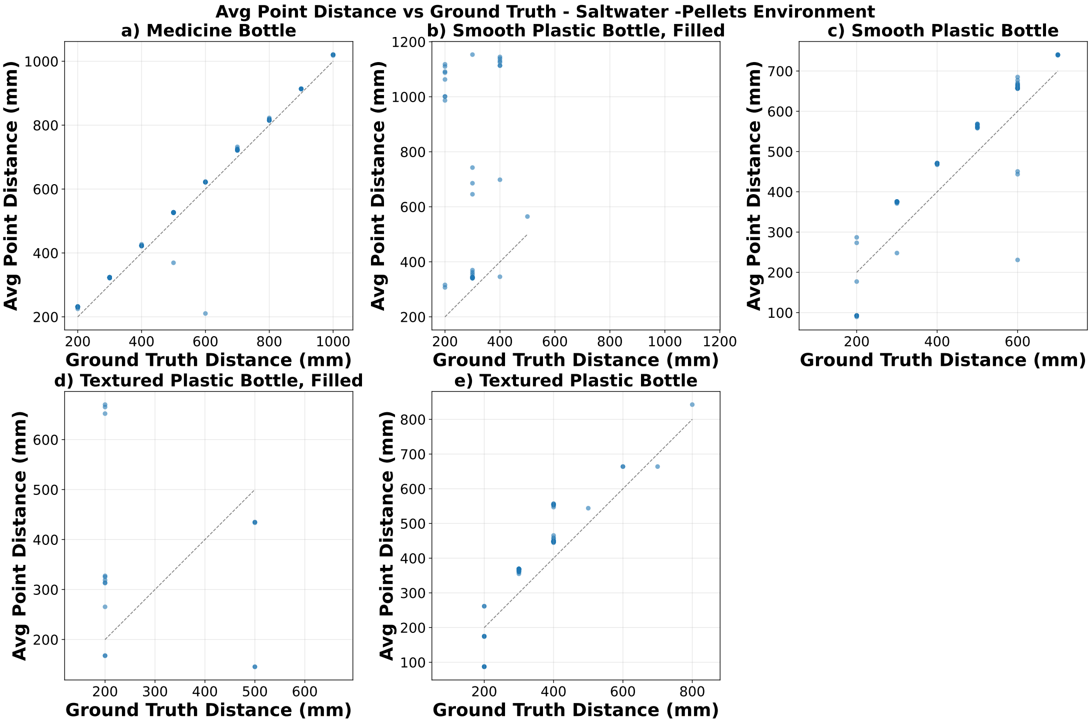


Figure 5.9: Point distance samples for Saltwater with Pellets environment. Suspended particulates produce the most challenging condition, with the largest errors and widest scatter across all bottle types. The filled smooth bottle is nearly undetectable or yields highly unreliable ranges. This represents the practical lower bound of the unmodified stereo pipeline.

but the all-detections RMSE is higher (445 mm) due to occasional invalid or zero returns. Textured bottles (empty and filled) achieve the tightest clustering (success-only RMSE: 31 mm and 17 mm, respectively), while the medicine bottle shows stable performance (success-only RMSE: 20 mm). This establishes the in-air ceiling for object-level ranging: high success rates, low RMSE, and minimal range-dependent bias.

Freshwater Day. Moving to Freshwater Day (Fig. 5.11), success rates decline and RMSE increases across all bottle types. For the filled smooth (transparent) bottle, success-only RMSE rises dramatically from 13 mm (Air) to 333 mm (Freshwater Day), reflecting refraction-induced depth ambiguity: the stereo matcher often locks onto the rear wall of the water-filled bottle or returns invalid (NaN) depth. All-detections RMSE for this bottle shows a small decrease (445→432 mm), but this is misleading—it reflects a higher proportion of invalid returns (which may be excluded from RMSE calculation if coded as NaN but included as zeros if the SDK returns zero). Success-only filtering exposes the true degradation. Textured bottles remain comparatively robust but still show elevated error: textured empty success-only RMSE rises from 31 mm to 154 mm, while smooth empty increases from 30 mm to 68 mm. The medicine bottle, consistent with point-distance trends, shows modest degradation (20→48 mm). Object-distance error grows with range, particularly underwater, indicating that backscatter and refraction compound over longer optical paths.

Freshwater Night. Freshwater Night conditions (Fig. 5.12) exacerbate the degradation observed in Freshwater Day. Success rates drop further, especially at longer ranges and for transparent bottles. The filled smooth bottle exhibits wider scatter and more frequent invalid returns, while textured bottles—though still comparatively stable—show increased RMSE. The range-dependent error trend steepens: at 1400 mm, even textured bottles yield occasional high-error outliers. Reduced lighting interacts with refraction and turbidity to reduce stereo correspondence quality, yielding fewer valid object distances and larger errors when ranges are returned. This underscores the compounding effect of environmental stressors on object-level ranging.

Freshwater with `basicEnhance`. Applying `basicEnhance` pre-processing (Fig. 5.13) increases object-distance success rates from 41.6% (baseline Freshwater) to 50.8%, indicating that enhanced edges and contrast stabilise detection presence. However, this comes at the cost of higher RMSE: object-distance RMSE rises from 146.2 mm to 209.3 mm (all detections). This trade-off is expected and mirrors the point-distance findings: because basicEnhance was designed to improve detection stability rather than ranging accuracy (Section 4.5.3), the enhancement amplifies both target and non-target structure, yielding more frequent valid detections but noisier depth estimates. For applications prioritising detection stability over ranging precision—such as object tracking or occupancy mapping—this trade-off may be acceptable; for manipulation or precise navigation, the elevated error is problematic.

Saltwater. Saltwater conditions (Fig. 5.14) introduce additional backscatter and turbidity relative to freshwater, further degrading object-distance reliability. Success rates decline, particularly for transparent and smooth bottles, and RMSE increases across all types. The filled smooth bottle remains the dominant error source, with frequent invalid or highly erroneous ranges. Textured bottles maintain partial robustness but show wider error distributions and more frequent outliers at longer ranges. The medicine bottle continues to perform comparatively well, suggesting that small, high-contrast, textured targets are more resilient to increased turbidity. Overall, Saltwater represents a significant step down in ranging reliability relative to Freshwater Day, with the gap widening as range increases.

Saltwater with Pellets. The most challenging condition—Saltwater with suspended Pellets (Fig. 5.15)—produces the lowest success rates and highest RMSE across all bottle types. Suspended particulates increase backscatter, reduce contrast, and introduce spurious stereo correspondences, causing the depth estimator to fail frequently or return highly erroneous ranges. The filled smooth bottle is nearly unusable for ranging: valid returns are rare, and when present, errors are extreme. Even textured bottles—previously the most robust—show substantially degraded performance, with success rates dropping and error distributions widening. This condition represents the practical lower bound of the unmodified stereo pipeline: object-level ranging becomes unreliable for all but the highest-contrast, most heavily textured targets, and even then only at shorter ranges.

Filtering and practical implications. Across all environments, success-only filtering (restricting analysis to detections that return valid, non-zero ranges) reduces absolute RMSE values and exposes the true magnitude of underwater degradation. However, filtering also discards data, reducing coverage. A practical deployment strategy would combine success-only filtering with confidence-gated pooling (using the ZED SDK's per-pixel confidence map to suppress low-quality depth within the detection region) and temporal smoothing (e.g., exponentially weighted moving average, EWMA) to stabilise frame-to-frame estimates. Even with these heuristics, the air–water gap remains substantial for refractive cases (filled transparent bottles), indicating that closing the gap will require geometry-aware methods that explicitly model refraction and 3D constraints.

Takeaway. Object-distance RMSE and success rates systematically degrade as environmental complexity increases (Air $<$ Freshwater Day $<$ Freshwater Night $<$ Saltwater $<$ Saltwater+Pellets). The filled transparent bottle is consistently the hardest target, with success-only RMSE rising from 13 mm (Air) to 333 mm (Freshwater Day) and beyond. Textured surfaces and the small medicine bottle are comparatively robust but still show substantial underwater degradation. Enhancement (`basicEnhance`) increases success rates at the cost of higher RMSE. For reliable underwater ranging, especially of transparent objects, methods must go beyond drop-in pre-processing and address refraction, backscatter, and correspondence ambiguity explicitly.

5.1.6 Classifier Score–Distance Relationship

Detector confidence—the YOLOv8 model's predicted probability that a detection corresponds to the target class (`bottle`)—provides a complementary quality signal to range estimates. Confidence reflects the model's certainty given the visual appearance of the target: high confidence

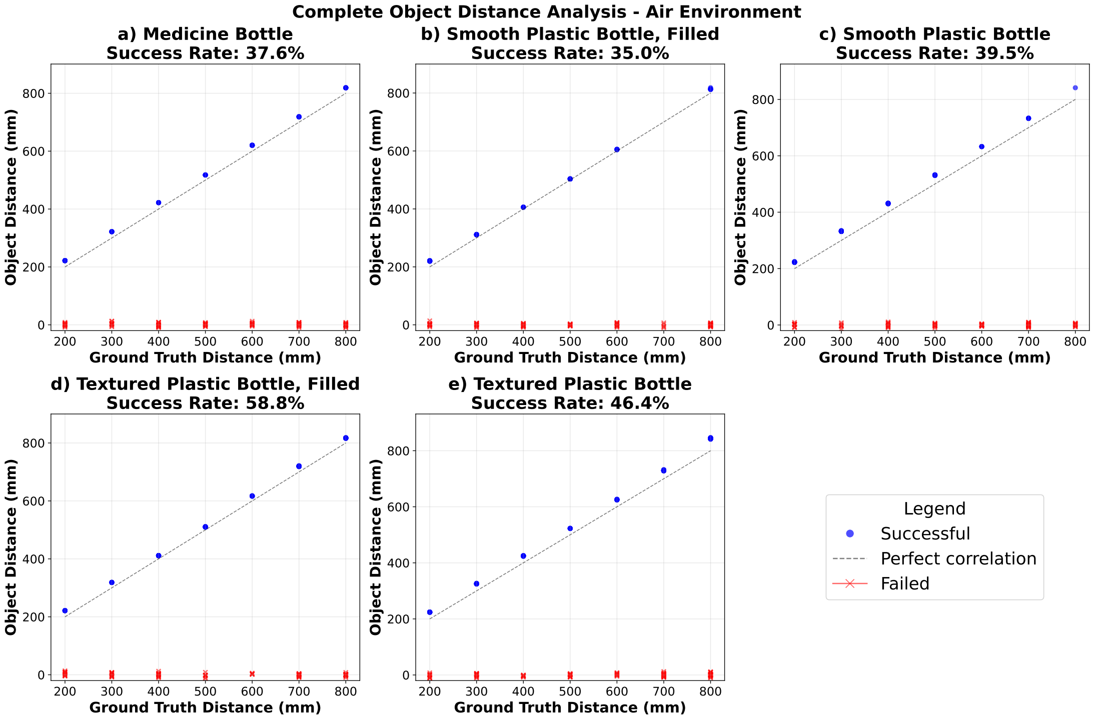


Figure 5.10: Object distance samples for Air environment, including failed detections and overall success rates. In-air performance is excellent: high success rates (>95%), tight clustering around ground truth, and low success-only RMSE across all bottle types. The filled smooth (transparent) bottle shows modest scatter (success-only RMSE: 13 mm); textured bottles achieve the tightest clustering (success-only RMSE: 17–31 mm).

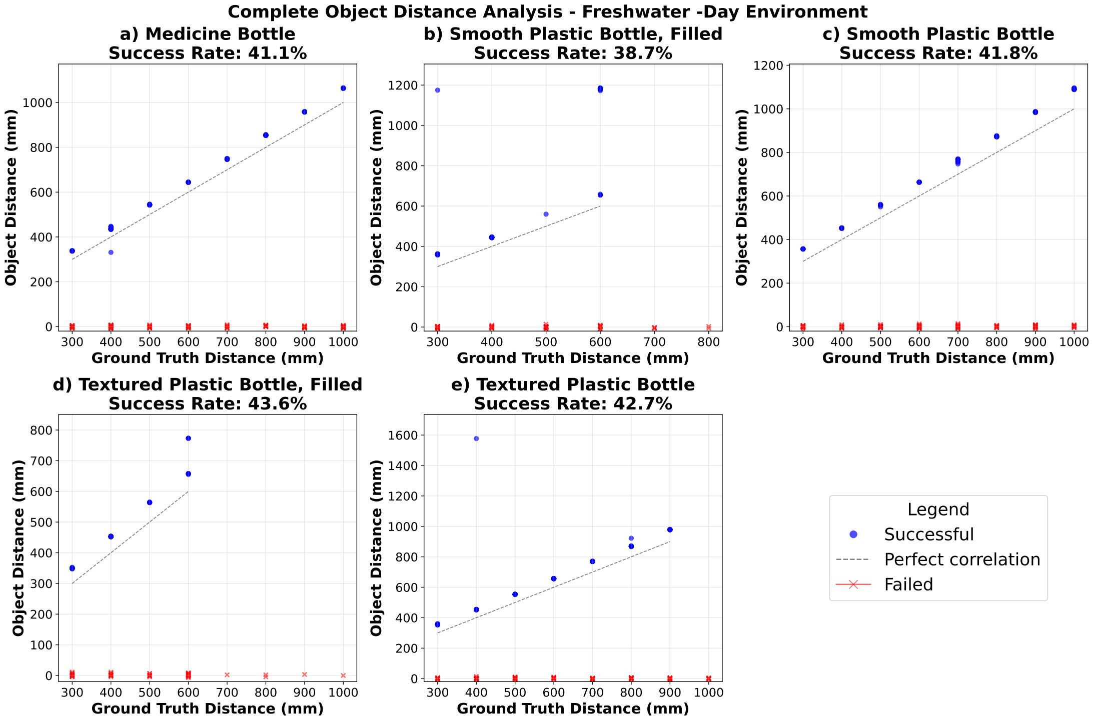


Figure 5.11: Object distance samples for Freshwater Day environment. Success rates decline and RMSE increases across all bottle types. The filled smooth (transparent) bottle exhibits dramatic success-only RMSE increase (13→333 mm), reflecting refraction-induced depth ambiguity. Textured bottles remain comparatively robust but show elevated error (e.g., textured empty: 31→154 mm). Range-dependent error trends emerge, steepening at longer distances.

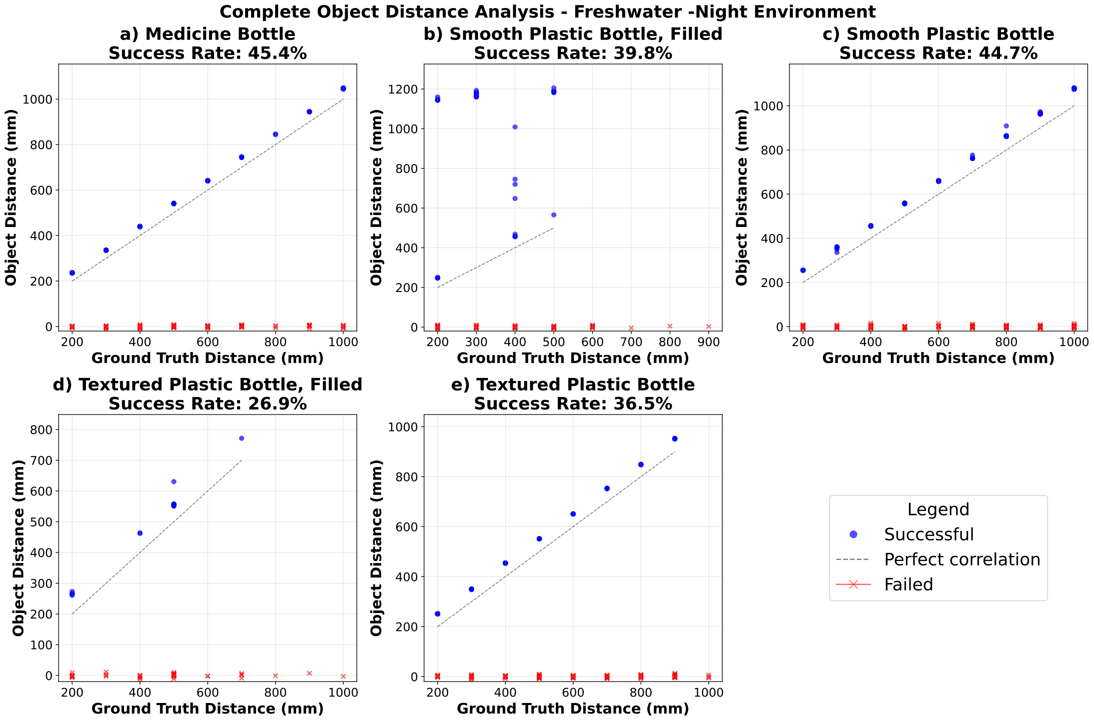


Figure 5.12: Object distance samples for Freshwater Night environment. Reduced lighting further degrades success rates and increases RMSE relative to Freshwater Day. The filled smooth bottle exhibits wider scatter and more frequent invalid returns; textured bottles show increased RMSE and steeper range-dependent error trends. The compounding effect of lighting and refraction reduces stereo correspondence quality.

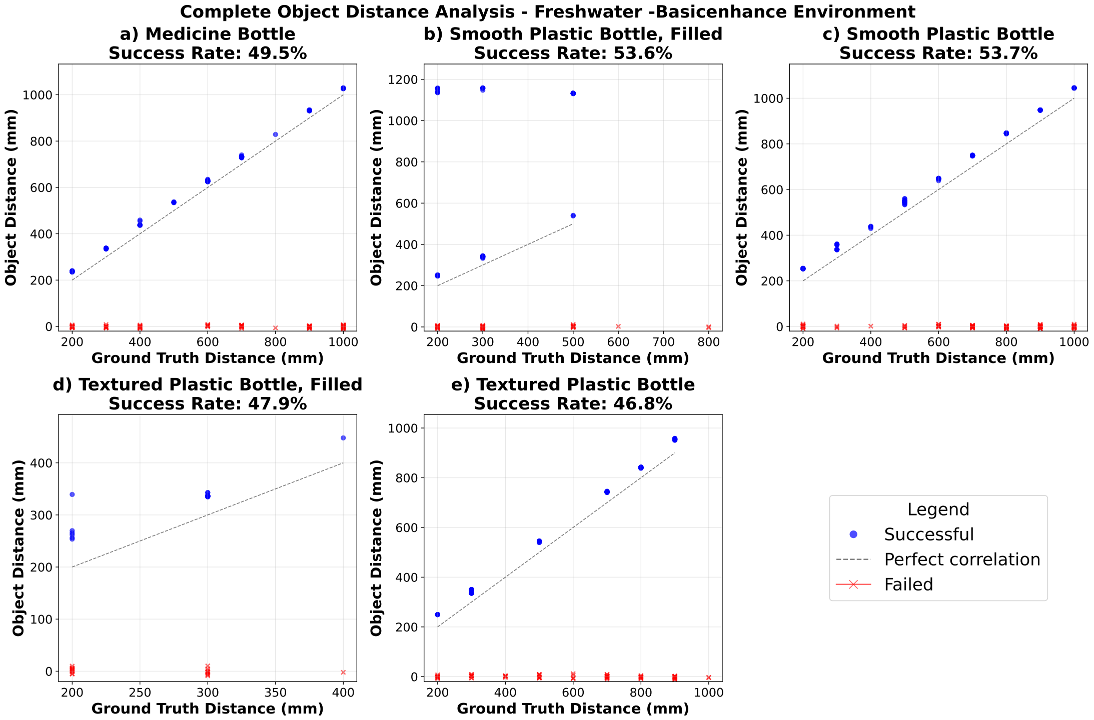


Figure 5.13: Object distance samples for Freshwater with `basicEnhance` pre-processing. Success rates improve (41.6%→50.8%), indicating more frequent valid detections, but RMSE increases (146.2→209.3 mm). This expected trade-off reflects the design intent: basicEnhance targets detection stability rather than ranging accuracy, and enhanced edges amplify both target and non-target structure.

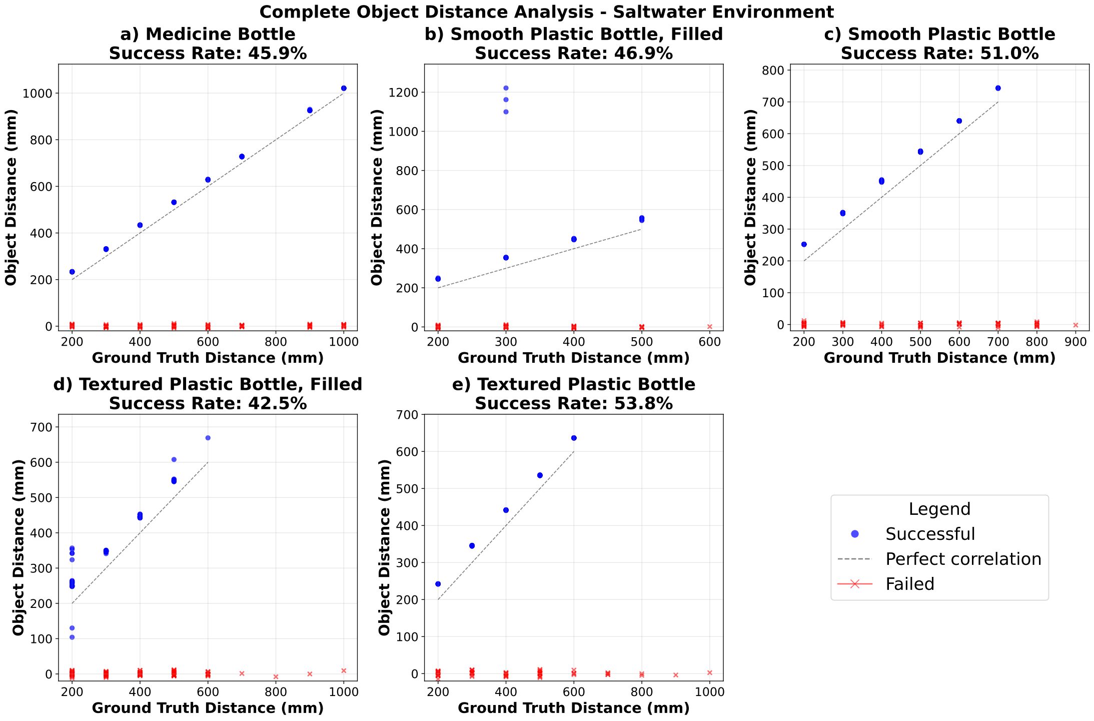


Figure 5.14: Object distance samples for Saltwater environment. Additional backscatter and turbidity relative to freshwater further degrade success rates and increase RMSE across all bottle types. The filled smooth bottle remains the dominant error source; textured bottles maintain partial robustness but show wider error distributions. Range-dependent error trends widen, particularly at longer distances.

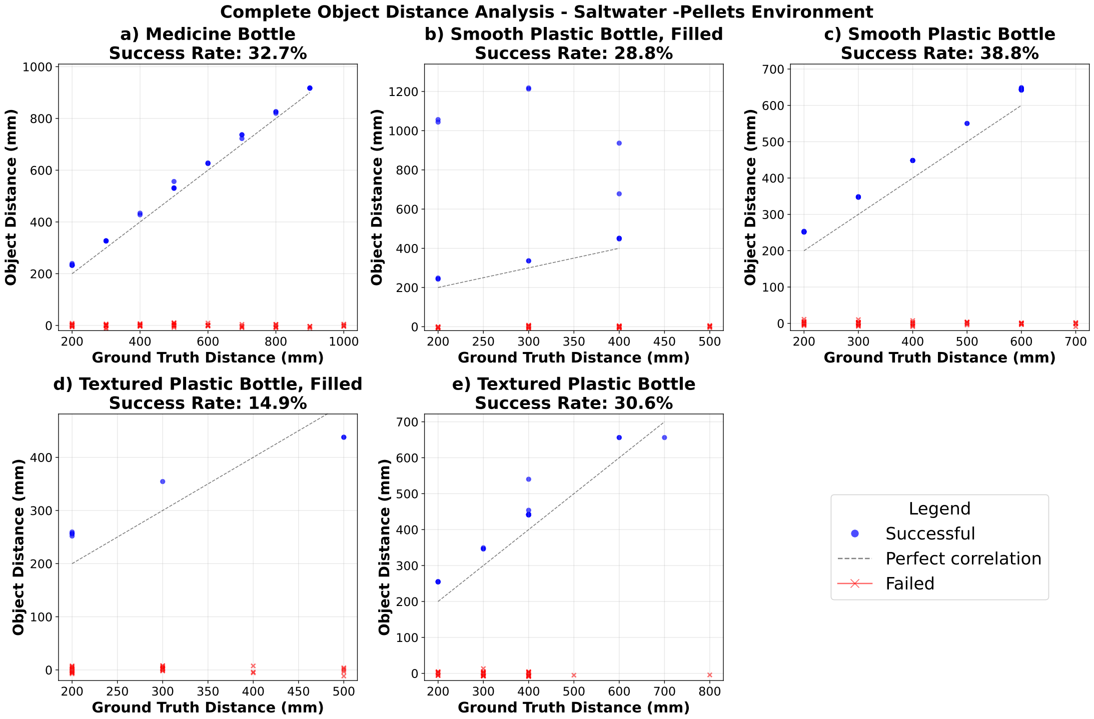


Figure 5.15: Object distance samples for Saltwater with Pellets environment. Suspended particulates produce the lowest success rates and highest RMSE across all bottle types. The filled smooth bottle is nearly unusable for ranging; even textured bottles show substantially degraded performance. This represents the practical lower bound of the unmodified stereo pipeline: object-level ranging becomes unreliable for all but the highest-contrast targets at short ranges.

indicates clear, unambiguous features (edges, texture, shape), while low confidence suggests degraded or ambiguous inputs. By examining how confidence varies with distance and environment, we can assess whether confidence-gating (filtering detections below a threshold) might stabilise ranging without retraining the detector. Figures 5.16–5.21 present detector confidence versus ground-truth distance for all six environments, with each panel separated by bottle type to reveal per-target trends.

Air baseline. In Air (Fig. 5.16), detector confidence remains consistently high across all bottle types and distances (400–1400 mm). Textured bottles (empty and filled) yield confidence values predominantly in the 70–85% range, with minimal scatter and no systematic distance-dependent decline. Smooth bottles (empty and filled transparent) show slightly wider confidence distributions but still cluster in the 60–80% range. The medicine bottle achieves similarly stable confidence (75–85%). Classification accuracy is near-perfect: smooth empty and textured empty achieve 100%, while filled smooth reaches 98.8%. Mean confidence (all detections) ranges from 71.5% (smooth empty) to 79.8% (medicine), and success-only filtering (restricting to correct `bottle` classifications) yields marginally higher values. This establishes the in-air baseline: high, stable confidence across the full operational range, with minimal need for confidence-gating.

Freshwater Day. Moving to Freshwater Day (Fig. 5.17), confidence levels shift downward and widen for all bottle types. Classification accuracy drops: smooth empty falls to 89.7%, filled smooth to 87.7%, and textured empty to 74.4%. Medicine and textured filled remain at 100% accuracy but with reduced confidence. Mean confidence (all detections) decreases systematically: 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%. Success-only filtering raises confidence modestly (e.g., textured empty: 59.1→62.0%), but the underwater penalty persists. The confidence–distance scatter plots reveal wider tails extending to low confidence ($<$50%) at all ranges, most prominently for textured empty and filled smooth bottles. This suggests that refraction, reduced contrast, and backscatter degrade feature quality, reducing the model's certainty even when detections are geometrically correct. A modest confidence threshold (e.g., 55–60%) would filter the lowest-quality detections while retaining the majority, potentially stabilising range estimates at the cost of reduced coverage.

Freshwater Night. Freshwater Night conditions (Fig. 5.18) exacerbate the confidence degradation observed in Freshwater Day. Confidence distributions widen further, with more frequent low-confidence detections across all bottle types. The distance-dependent decline in confidence becomes more pronounced: at longer ranges (1000–1400 mm), even textured bottles—previously the most robust—show confidence values dropping below 50%. Classification accuracy declines for most types, and mean confidence decreases across the board. The low-confidence tail extends deeper, indicating that reduced lighting compounds refraction and turbidity effects, making target features less distinctive. Confidence-gating becomes more critical in this regime: a threshold of around 60% would discard a substantial fraction of detections but retain those most likely to yield reliable ranges. However, this trades coverage for quality; a practical deployment must balance these competing objectives based on task requirements (e.g., manipulation vs occupancy mapping).

Freshwater with `basicEnhance`. Applying `basicEnhance` pre-processing (Fig. 5.19) yields a mixed outcome for classifier confidence. Classification accuracy drops slightly (90.3%→ 87.6%), indicating that enhanced edges occasionally produce false positives or confuse the detector. Mean confidence trends are variable: some bottle types show modest increases, others show decreases, but overall the enhancement does not systematically improve confidence levels. The confidence–distance scatter plots reveal that while the low-confidence tail is reduced slightly (consistent with increased success rates), the high-confidence cluster is not substantially elevated. This aligns with earlier findings that `basicEnhance` stabilises detection presence but does not improve feature quality for ranging or classification certainty. For confidence-gated applications, `basicEnhance` offers marginal benefit and may be best combined with other techniques (e.g., CLAHE, Grey-World colour normalisation) to address lighting and contrast more systematically.

Saltwater. Saltwater conditions (Fig. 5.20) further degrade detector confidence relative to freshwater. Confidence distributions shift downward and widen, particularly for transparent and smooth bottles. Classification accuracy declines for most types, and mean confidence drops across the board. The distance-dependent decline steepens: at 1400 mm, confidence values frequently fall below 50%, even for textured bottles. The low-confidence tail extends deeper, indicating that increased backscatter and turbidity reduce feature distinctiveness, making the model less certain of its predictions. Confidence-gating at around 60% would discard a large fraction of detections, especially at longer ranges, but would retain those most likely to correspond to valid targets. The filled smooth (transparent) bottle exhibits the widest confidence distribution and lowest mean confidence, consistent with its status as the hardest target across all metrics (point distance, object distance, classification accuracy).

Saltwater with Pellets. The most challenging condition—Saltwater with suspended Pellets (Fig. 5.21)—produces the lowest confidence levels and widest scatter across all bottle types. Classification accuracy drops substantially for most types, and mean confidence decreases to the lowest observed values. The confidence–distance scatter plots reveal frequent low-confidence detections ($<$40%) at all ranges, with even high-contrast textured bottles showing reduced certainty. The distance-dependent decline is extreme: beyond 1000 mm, confidence values are highly variable and often fall below 50%, indicating that suspended particulates obscure target features and introduce spurious edges that confuse the detector. Confidence-gating at even modest thresholds (e.g., 55%) would discard the majority of detections, leaving only short-range, high-contrast targets. This represents the practical lower bound of the unmodified detector's utility: confidence levels are too low and too variable to support reliable gating without substantial coverage loss.

Confidence-gating policy. Across all environments, detector confidence decreases with distance and shifts downward as environmental complexity increases (Air $<$ Freshwater Day $<$ Freshwater Night $<$ Saltwater $<$ Saltwater+Pellets). The filled smooth (transparent) bottle consistently exhibits the lowest and most variable confidence, aligning with its poor performance in point-distance and object-distance metrics. Textured bottles maintain comparatively higher confidence but still show substantial underwater degradation. A practical confidence-gating policy would set environment- and range-dependent thresholds: for Freshwater Day, a threshold of 55–60% would retain most detections while filtering low-quality outliers; for Saltwater+Pellets, even a 50% threshold would discard the majority of detections beyond 1000 mm. Gating can be combined with success-only filtering (restricting to detections that return valid ranges) and temporal smoothing (EWMA) to further stabilise estimates. However, confidence-gating alone does not close the air–water gap for refractive cases; it mitigates but does not eliminate the underlying depth-estimation and feature-quality challenges.

Takeaway. Detector confidence systematically declines underwater, with the largest drops for transparent and smooth bottles. Classification accuracy follows a similar trend, falling from near-perfect in air to 75–90% in freshwater and lower in saltwater. Mean confidence (both all-detections and success-only) decreases across environments, and the confidence–distance relationship reveals wider low-confidence tails underwater. Confidence-gating offers a lightweight mechanism to filter unreliable detections without retraining, but thresholds must be tuned per environment and range to balance coverage and quality. Enhancement (`basicEnhance`) does not systematically improve confidence and slightly reduces accuracy, indicating limited utility for classification robustness. For reliable underwater detection and ranging of transparent objects, confidence-gating should be combined with geometry-aware methods that address the root causes of feature degradation: refraction, backscatter, and reduced contrast.

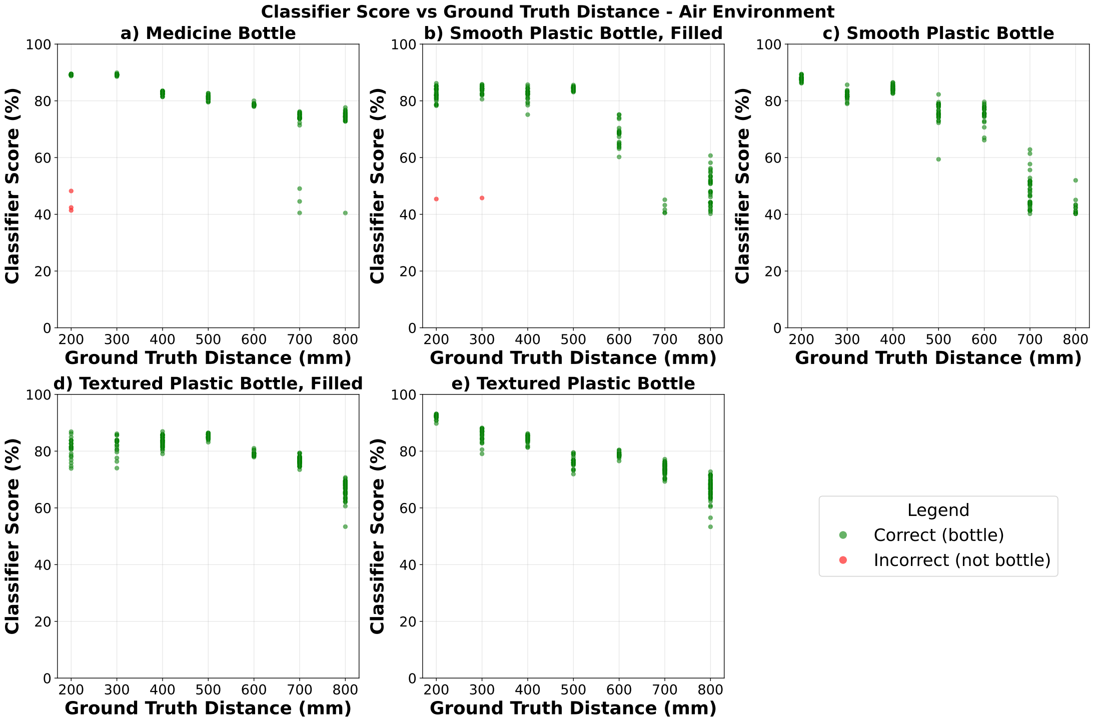


Figure 5.16: Detector confidence versus ground-truth distance for Air environment, separated by bottle type. Confidence remains consistently high (60–85%) across all types and distances, with minimal scatter. Classification accuracy is near-perfect (smooth/textured empty: 100%; filled smooth: 98.8%), and mean confidence ranges from 71.5% to 79.8%. This establishes the in-air baseline for detector certainty.

5.1.7 Object Dimension Accuracy

Beyond ranging, the ZED SDK provides per-detection 3D object dimensions (width, height, depth) computed from stereo depth within the detection bounding box. This subsection evaluates dimension accuracy against ground-truth measurements (Table 4.3) to characterise the stereo pipeline's object sizing capability—a task that faces similar challenges to ranging (refraction, low contrast, correspondence failures) but targets a distinct output modality.

Success rates. Of 4,407 total detections across all environments, 1,929 (43.8%) returned valid 3D dimensions (all three dimensions non-zero). This success rate closely matches the Ob-

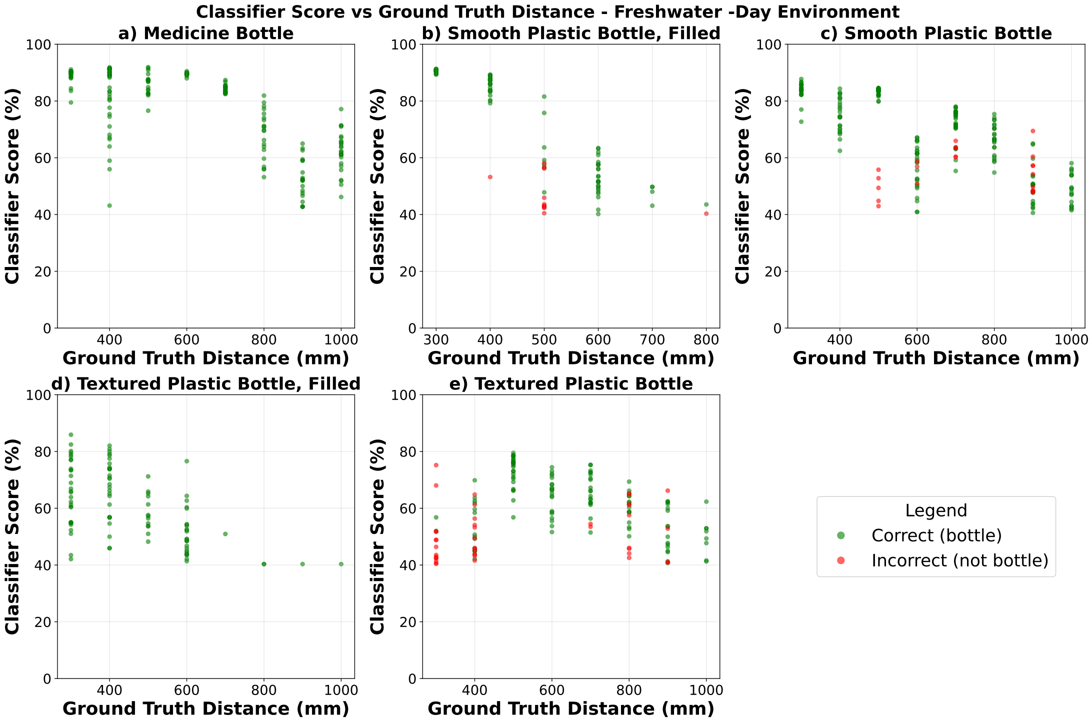


Figure 5.17: Detector confidence versus ground-truth distance for Freshwater Day environment. Confidence levels shift downward and widen relative to Air, with low-confidence tails (<50%) emerging for all bottle types. Classification accuracy drops (smooth empty: 89.7%; filled smooth: 87.7%; textured empty: 74.4%), and mean confidence decreases systematically. A modest confidence threshold (55–60%) would filter low-quality detections while retaining the majority.

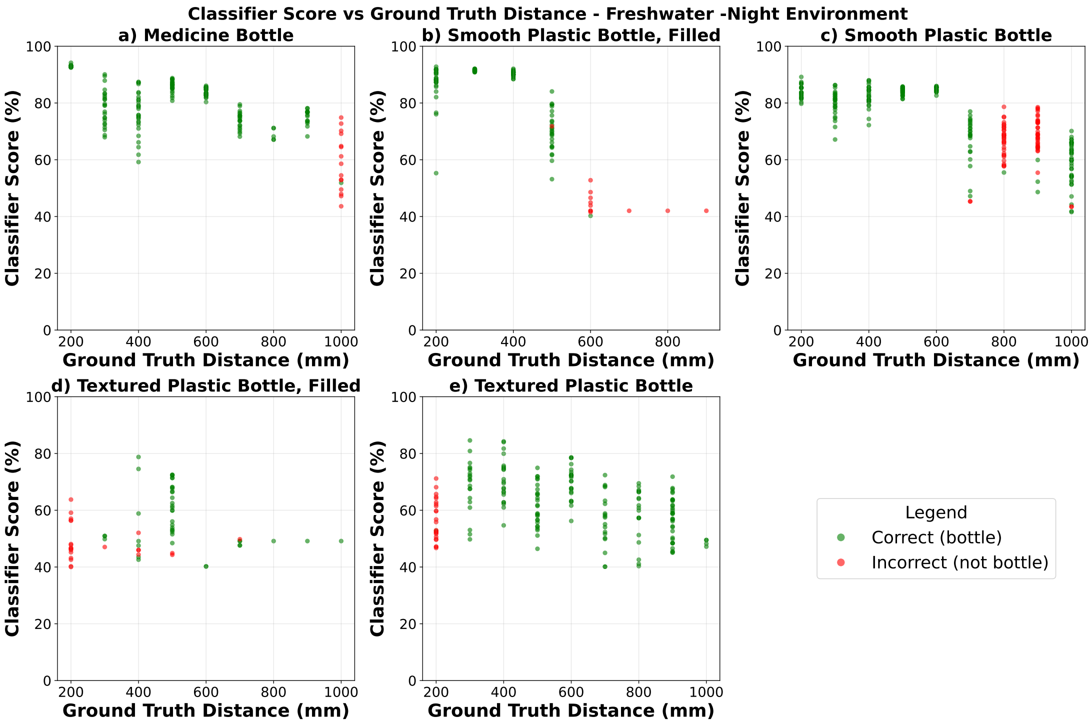


Figure 5.18: Detector confidence versus ground-truth distance for Freshwater Night environment. Reduced lighting further degrades confidence, with wider distributions and deeper low-confidence tails relative to Freshwater Day. Classification accuracy declines across most types, and the distance-dependent confidence drop becomes more pronounced. Confidence-gating becomes more critical but trades coverage for quality.

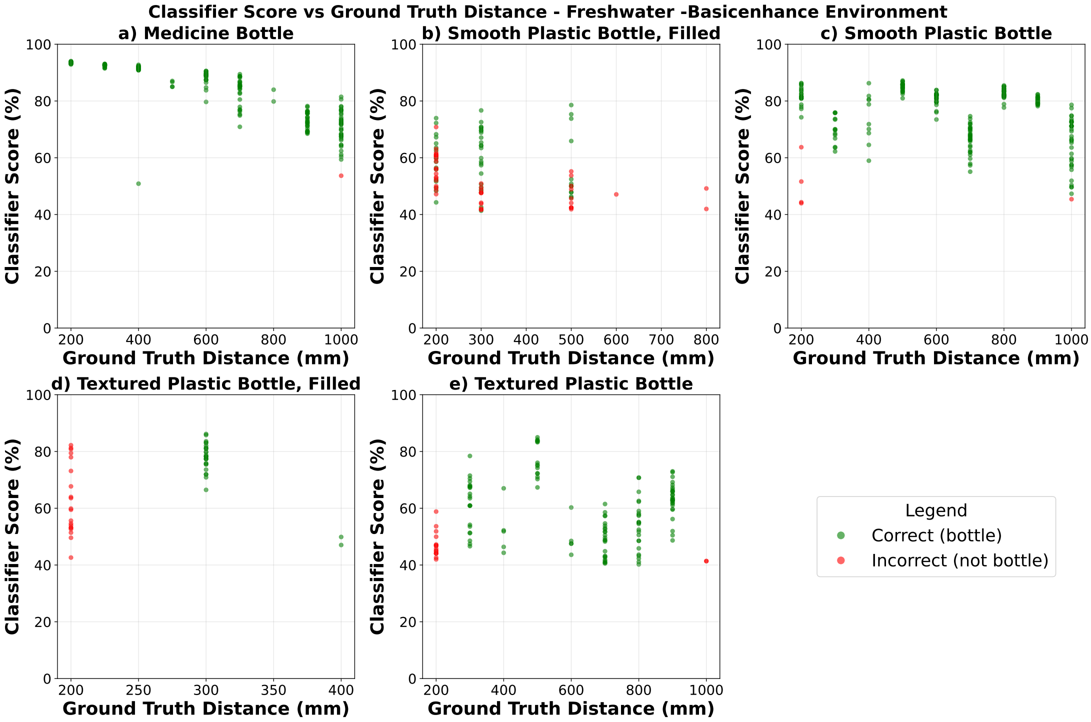


Figure 5.19: Detector confidence versus ground-truth distance for Freshwater with `basicEnhance` pre-processing. Classification accuracy drops slightly (90.3%→87.6%), and mean confidence trends are variable across bottle types. The low-confidence tail is reduced modestly, but high-confidence clusters are not substantially elevated. Enhancement stabilises detection presence but does not systematically improve confidence or classification certainty.

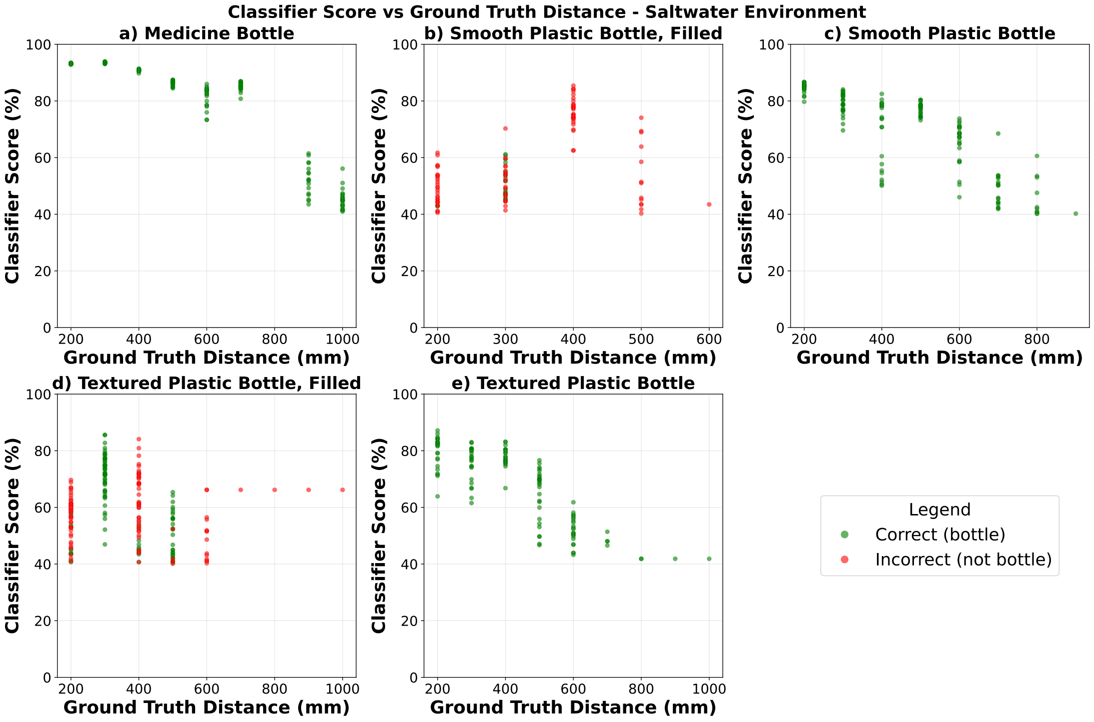


Figure 5.20: Detector confidence versus ground-truth distance for Saltwater environment. Confidence distributions shift downward and widen relative to freshwater, with steeper distance-dependent decline. Classification accuracy declines for most types, and mean confidence drops across the board. At 1400 mm, confidence frequently falls below 50%, even for textured bottles. The filled smooth (transparent) bottle exhibits the widest confidence distribution and lowest mean confidence.

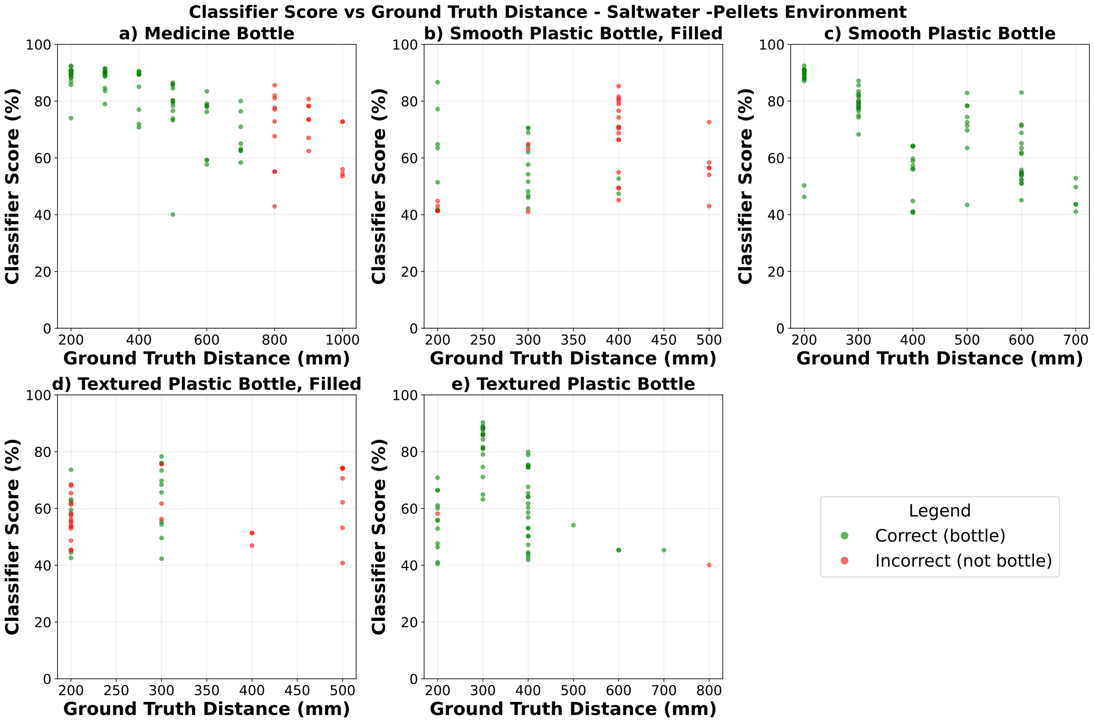


Figure 5.21: Detector confidence versus ground-truth distance for Saltwater with Pellets environment. Suspended particulates produce the lowest confidence levels and widest scatter across all bottle types. Classification accuracy drops substantially, and frequent low-confidence detections (<40%) occur at all ranges. Beyond 1000 mm, confidence is highly variable and often below 50%, indicating that particulates obscure target features. This represents the practical lower bound of the unmodified detector's utility for confidence-gated filtering.

ject Distance success rate (Table 5.1), indicating that dimension estimation and ranging face common underlying challenges. The environmental gradient mirrors distance findings: freshwater-basicEnhance achieves the highest dimension success rate (51.5%), while saltwater-pellets shows the lowest (33.1%).

Gross measurement errors. Preliminary analysis revealed that 214 of 1,929 successful measurements (11.1%) contain gross errors where one or more dimensions deviate by more than 2× from ground truth. These outliers arise from stereo matching failures—for example, width estimates of 2,208 mm for a 50 mm bottle—and would dominate summary statistics if included. Following standard practice, we report both raw metrics (all successful measurements) and filtered metrics (outliers removed) to expose the impact of gross errors.

Filtered accuracy. After removing gross outliers (n = 1,715 measurements), all three dimensions achieve reasonable accuracy, reported here as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE):

- **Width:** MAE 4.0 mm, MAPE 7.7%, Bias +3.4 mm
- **Height:** MAE 8.7 mm, MAPE 5.7%, Bias +4.8 mm
- **Depth:** MAE 4.0 mm, MAPE 7.7%, Bias +3.4 mm

All dimensions exhibit systematic positive bias (overestimation), consistent with the ranging bias observed in Section 5.1.1. Height shows the best percentage accuracy (5.7% MAPE) despite having the largest absolute error (8.7 mm MAE), reflecting the larger ground-truth values (118–170 mm vs 50–55 mm for width/depth).

Raw vs filtered metrics. The contrast between raw and filtered metrics illustrates the impact of gross outliers:

- **Width:** Raw MAPE 181.1% vs Filtered MAPE 7.7%
- **Height:** Raw MAPE 45.0% vs Filtered MAPE 5.7%
- **Depth:** Raw MAPE 19.6% vs Filtered MAPE 7.7%

Depth shows the smallest raw MAPE (19.6%), suggesting that depth-aligned measurements (along the camera optical axis) are more robust to stereo matching errors than perpendicular measurements (width/height).

Cross-environment patterns. Figure 5.22 presents RMSE heatmaps for each dimension across environments and bottle types (outliers removed). The environmental gradient mirrors distance findings: Air achieves the lowest RMSE across all dimensions and bottle types, while underwater conditions show elevated error. The filled smooth (transparent) bottle—identified as the hardest target for ranging—also exhibits elevated dimension RMSE in several conditions. Height RMSE is generally larger than Width or Depth in absolute terms, but this reflects the larger GT values (118–170 mm) rather than worse relative accuracy.

Implications. The dimension analysis reveals that the ZED SDK can estimate object sizes with 5–8% accuracy when successful and non-outlier, but produces gross measurement errors in approximately 11% of cases. This is sufficient for coarse object sizing or classification by size category, but applications requiring precise dimensions should implement outlier detection (e.g., temporal consistency checks, geometric constraints). The similar success rates for dimensions and ranging suggest that improvements to stereo correspondence quality would benefit both tasks. Full methodology, additional visualisations (including raw RMSE with outliers and success rate heatmaps), and detailed metrics by environment and bottle type are provided in Appendix B.

5.2 Chapter summary

This chapter presented a comprehensive cross-environment analysis of detection and ranging performance for transparent plastic bottles using the UW-TransStereo dataset and benchmark introduced in Chapter 4. The results established clear environmental gradients: performance degrades systematically from Air through Freshwater Day and Night to Saltwater and Saltwater with Pellets, with RMSE increasing, success rates falling, and detector confidence declining as conditions become more challenging. Surface texture emerged as the dominant factor determining robustness: textured bottles maintained comparatively reliable detection and ranging even in degraded conditions, while the filled smooth transparent bottle consistently exhibited the highest error rates and lowest success rates across all environments.

The refractive failure indicators introduced in Chapter 2—Rear-Wall Bias (RWB) and Bias—were computed for all environments (Table 5.1). RWB directly quantified the index-matching failure mode: in Air, RWB was exactly 0% as expected (no refractive medium to cause optical transparency), while the filled smooth (water-filled transparent) bottle exhibited extreme RWB underwater, reaching 55.6% in Freshwater Night—meaning more than half of valid depth estimates locked onto the tank wall rather than the bottle surface. Bias correlated with RWB across environments: Freshwater Night showed both the highest RWB (7.4%) and the largest

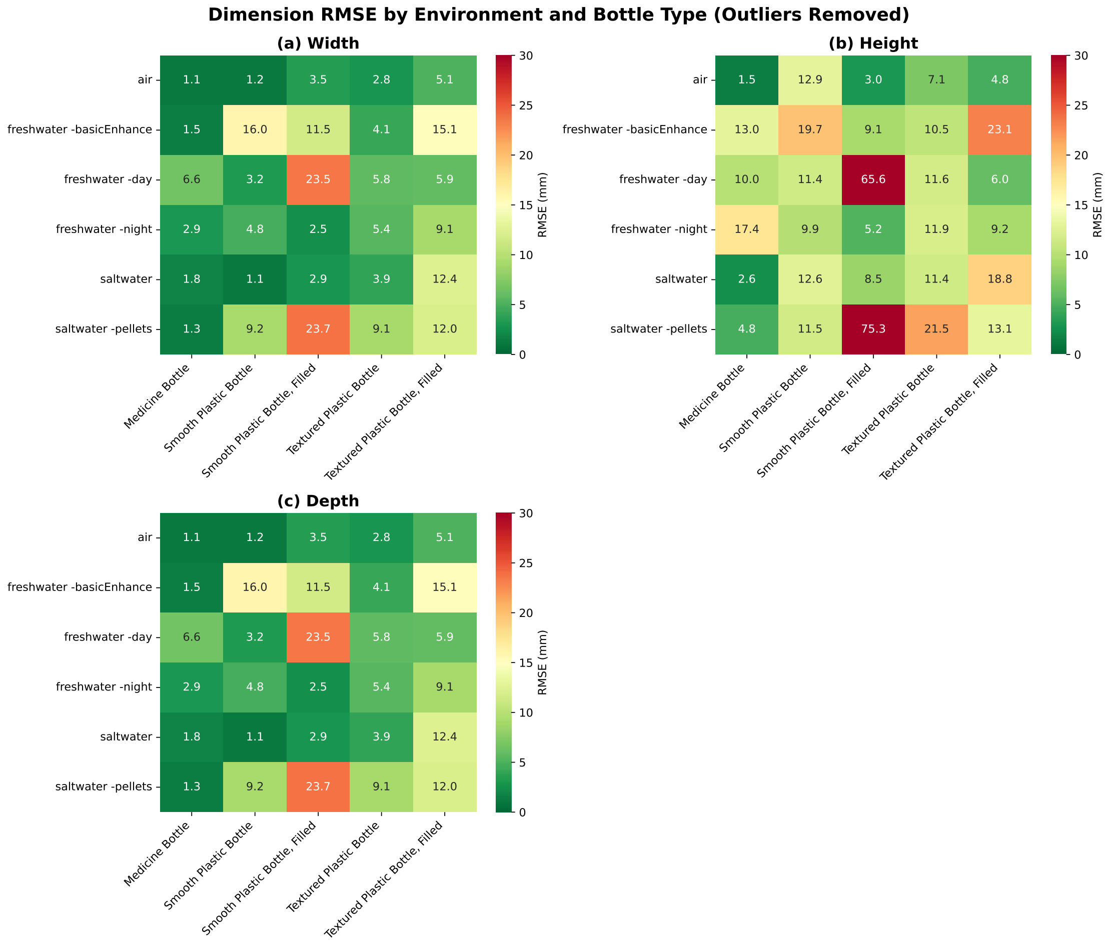


Figure 5.22: Dimension RMSE by environment and bottle type (outliers removed, $n = 1{,}715$ measurements). Left: Width RMSE. Centre: Height RMSE. Right: Depth RMSE. Darker/redder cells indicate higher error. Air achieves the lowest RMSE across all dimensions; the environmental gradient mirrors distance findings. Height shows larger absolute RMSE than Width/Depth due to larger GT values (118–170 mm vs 50–55 mm).

Bias (126 mm), confirming that rear-wall locking drives systematic range overestimation. Interestingly, Saltwater conditions exhibited lower RWB than Freshwater despite greater turbidity, suggesting that backscatter may provide weak texture cues that help anchor disparity to nearer surfaces. These findings validate the evaluation protocol and demonstrate that RWB and Bias provide actionable diagnostics for operator-aid gating.

The comparison of manual point distance and automatic object distance revealed that both estimators capture the same qualitative trends and rank bottle types identically, but differ in absolute error magnitude and failure modes. Object distance typically yields higher RMSE and more frequent invalid returns than point distance, reflecting the SDK's difficulty integrating depth over large refractive or low-contrast regions. Success-only filtering exposed the true magnitude of underwater degradation by excluding invalid returns, showing that when the system does produce a range estimate for transparent bottles underwater, the error can be substantial (e.g., success-only object-distance RMSE rising from 13 mm in Air to 333 mm in Freshwater Day for the filled smooth bottle).

Statistical tests confirmed that, for this experimental setup, saltwater conditions provide modest but significant advantages over Freshwater Day for both point-distance and object-distance accuracy and for object-distance success rates, despite greater nominal turbidity. This counterintuitive result suggests that the relationship between medium properties and stereo performance is complex and may depend on factors such as particulate size distribution, refractive index contrast, and lighting geometry. Detector confidence, by contrast, remained statistically indistinguishable between saltwater and freshwater, indicating that the performance gains are primarily in depth estimation rather than object recognition.

Distributional analysis confirmed that 96% of error distributions are non-normal, exhibiting heavy tails and extreme skewness (up to ±8) consistent with stereo matching failures; this validates the non-parametric statistical approach used throughout the chapter. Bias heatmaps revealed that the filled smooth bottle exhibits systematic overestimation of 400–600 cm in underwater environments, directly quantifying the magnitude of rear-wall locking beyond the RWB percentage metric. The full error distributions (Appendix A) revealed bimodal patterns for the filled smooth bottle, with distinct modes corresponding to successful surface detection and rear-wall locking, providing a visual characterisation of the index-matching failure mode.

The illustrative pre-processing variant, `basicEnhance`, demonstrated a fundamental stability–precision trade-off: it increased object-distance success rates by enhancing edges and contrast, but also raised RMSE and slightly reduced classification accuracy. This indicates that drop-in image enhancement alone cannot close the air–water gap for refractive cases; instead, methods must explicitly model refraction, geometry, and correspondence ambiguity to achieve robust underwater stereo ranging of transparent objects. Confidence-gating offers a lightweight deployment strategy to filter unreliable detections, but thresholds must be environment- and range-dependent, and even aggressive gating cannot fully compensate for the underlying depth-estimation challenges.

Beyond ranging, the analysis of object dimension accuracy (Section 5.1.7) revealed that the ZED SDK's 3D sizing capability faces similar challenges: dimension success rates (43.8%) closely match object-distance success rates, and the environmental gradient mirrors distance findings. When outliers are removed, dimension estimates achieve 5–8% MAPE—reasonable for coarse object sizing—but approximately 11% of successful measurements contain gross errors where dimensions deviate by more than 2× from ground truth. This outlier rate represents an important caveat for applications relying on stereo-derived object dimensions underwater.

Taken together, the results quantify the limits of commodity stereo pipelines for underwater transparent-object perception and motivate the development of refraction-aware, geometry-informed methods that can exploit multi-view constraints and learned priors to recover reliable depth in challenging refractive scenarios. Chapter 6 reflects on these findings in the broader context of stereo-informed telepresence and underwater manipulation, and outlines directions for future work.

Chapter 6

Conclusion

In this chapter, we summarise the main contributions of the thesis, outline key limitations across the presented systems and datasets, and propose directions for future work. We separate *summary of achievements*, *limitations*, and *future work* for clarity.

6.1 Summary of achievements

The thesis investigated stereo vision for augmented telepresence and underwater object ranging, culminating in the design and release of a multi-environment underwater stereo dataset and benchmark for transparent bottles. The main achievements are summarised below.

- **Stereo-informed augmented telepresence with MR overlays.** In Chapter 3, we implemented a stereo vision pipeline for augmented telepresence using commodity cameras and an MR interface. An initial unsynchronised OpenCV → Unity pipeline with dual webcams, disparity estimation (StereoBM), and UDP-based streaming was developed to provide stereo-derived distance cues overlaid on a 3D PluraView display, integrated with a haptic device and Unity-based manipulator model (Figs. 3.2 and 3.3) [33, 77, 1, 65]. Calibration and validation in air showed that the system could estimate distances across roughly 45–130 cm with an average RMSE of 9.13 cm; opaque and textured objects achieved low per-object RMSE (1.01–2.50 cm for the white bottle, metal bottle, teddy toy, and 3D-printed part), while a clear bottle was substantially harder (16.34 cm and occasional null readings) (Fig. 3.6) [33, 77, 1]. These experiments established a working stereo → MR distance overlay and highlighted transparent-object failure modes even in air.

- **Transition to a synchronised stereo front-end and underwater experiments.** The unsynchronised pipeline exposed critical blockers for real teleoperation—transport jitter, asynchronous capture vs rendering, and a monocular AR plugin—which motivated a move to a fully synchronised ZED Mini (ZEDM) front-end (Fig. 3.7) [33, 77, 1]. In the second part of Chapter 3, we integrated the ZED SDK with Unity (Fig. 3.8) and constructed an underwater test tank with a Perspex window, ZEDM mounting, and controllable WRGB lighting (Fig. 3.11) [33, 77]. A materials × environment study (air, underwater uncalibrated, underwater recalibrated) and a lighting study (smooth clear bottle under six colours in air) quantified how medium and illumination affect stereo ranging (Figs. 3.13–3.16) [33, 77, 1]. Recalibration markedly improved RMSE for opaque and textured materials underwater (e.g., white bottle 8.8 cm in air vs 18.5 cm underwater recalibrated), but filled transparent bottles remained unreliable (327–340 cm RMSE), and blue light caused catastrophic failures (585.3 cm RMSE) even in air. These results showed that while a vendor SDK and recalibration stabilise many cases, water-filled transparent plastics remain fundamentally difficult for standard stereo pipelines (Fig. 3.17) [33, 77, 1].

- **Design and release of the UW-TransStereo underwater stereo dataset.** Building on insights from Chapter 3, Chapter 4 introduced *UW-TransStereo*, a multi-environment underwater stereo dataset and benchmark focused on transparent bottle detection and ranging [74]. The dataset comprises 25 paired recordings (50 core files) across Air, Freshwater, Freshwater with an inference-time enhancement (`basicEnhance`), Saltwater, and Saltwater with Pellets, plus 216 qualitative screenshots; raw stereo is stored as `.svo2` (HD2K, 15 FPS, lossless), detector outputs as one `.xlsx` per recording, and screenshots as `.png` images [74]. Filenames follow a flat convention `[Environment]_[BottleType]_[filekind]_[timestamp].[ext]`, pairing SVO, detection spreadsheet, and screenshots, and a collated CSV provides >5,000 labelled measurements (environment, bottle type, ground-truth range, three vertical point-distance samples, object distance, 3D size, confidence, timestamps) with a documented schema and manifest (Tables 4.1 and 4.2) [74]. The dataset is released on Zenodo with an accompanying code repository and scripts to regenerate all reported metrics and plots [74].

- **Multi-environment benchmark with dual range estimators and cross-environment anal-**

ysis. Chapter 4 formalised two complementary range estimators—manual Average Point Distance (mean of three vertical samples per detection) and automatic Object Distance (ZED SDK per-object 3D centroid)—and an evaluation protocol combining RMSE, Bias, success rates, classification accuracy, mean confidence, and Rear-Wall Bias (RWB) to diagnose index-matching failure modes, reported under all-detections and success-only regimes [74, 74]. Using these, we established a six-condition benchmark (Air, Freshwater Day, Freshwater Night, Freshwater+`basicEnhance`, Saltwater, Saltwater+Pellets) with cross-environment heatmaps and per-environment scatter plots (Fig. 5.1 and Figs. 5.4–5.21) [74]. The analysis showed a clear environmental gradient (Air best, Saltwater+Pellets worst), with textured bottles remaining comparatively robust and the filled smooth (transparent) bottle consistently dominating error and failure rates [74]. Average Point Distance and Object Distance agreed on relative difficulty but differed in absolute error and failure modes, underscoring the importance of tracking both success probability and conditional accuracy [74, 74].

- **Quantitative comparison of freshwater and saltwater and practical guidance.** A formal statistical comparison in Chapter 4 showed that, for this setup, saltwater provides a modest but statistically significant advantage over Freshwater Day for both point-distance and object-distance MAE and for object-distance success rates, while classifier score distributions remain similar (Section 5.1.2). Specifically, saltwater improves average point-distance MAE by 58.7 cm (Cohen's $d = 0.289$, $p < 0.001$), object-distance MAE by 31.2 cm (Cohen's $d = 0.316$, $p < 0.001$), and increases object-distance success rates by 5.8% (47.4% vs 41.6%, $p = 0.023$), with no significant difference in detector confidence ($p = 0.598$) (Section 5.1.2). Distributional analysis of ranging errors validated the non-parametric statistical methodology by confirming that 96% of error distributions are non-normal, and revealed bimodal error patterns for water-filled transparent bottles that directly visualise the index-matching failure mode. Chapter 4 also evaluated a simple pre-processing variant (`basicEnhance`) and lightweight inference-time strategies (success-only filtering, confidence gating, temporal smoothing), concluding that enhancement improves success rates but raises RMSE and slightly reduces classification accuracy, and that environment- and range-specific confidence thresholds (e.g., 55–60% in Freshwater Day) can stabilise performance but do not close the air–water gap for refractive cases [74, 74].

6.2 Limitations

The work presented in this thesis is subject to several limitations at the system, dataset, and methodological levels. These limitations inform the future work proposed in Section 6.3.

- **Lab-scale rigs and environments.** All experiments were conducted in controlled lab settings with synthetic or tank-based environments. The unsynchronised stereo pipeline used a pair of webcams and a fixed baseline close to the human interpupillary distance [33, 77, 1], while the synchronised pipeline and UW-TransStereo dataset rely on a single ZED Mini camera mounted normal to a Perspex window in a bench-top tank with discrete media (Air, Freshwater, Saltwater, Saltwater+Pellets) and fixed distance markers (400–1400 mm in UW-TransStereo; 300–1000 cm in earlier underwater experiments) [74, 33]. This enables tight control and repeatability but does not capture the variability of open-water deployments (e.g., changing camera poses, vehicle motion, non-planar backgrounds, waves).
- **Restricted object and task diversity.** The telepresence experiments in Chapter 3 used a small set of objects (box, bottles, teddy, 3D-printed part, stone) for air validation and underwater materials × environment studies (white bottle, textured/smooth clear bottles and filled variants) [33, 77]. UW-TransStereo focuses on five bottle types (medicine, smooth empty, smooth filled, textured empty, textured filled) representative of common small plastic debris [74]. This object set is sufficient to reveal strong material- and texture-dependent effects but does not include larger containers, irregular shapes, bags, or non-bottle transparent objects, nor does it address tasks beyond detection and ranging (e.g., pose estimation or grasp planning).
- **Single stereo stack and limited algorithmic baselines.** All underwater experiments use the ZED Mini with the vendor's NEURAL depth mode and SDK object tracking, combined with a single modern detector (YOLOv8) in the dataset work [74, 74]. No systematic comparison is made against alternative stereo methods, depth configurations, or detectors, and no underwater-specific fine-tuning or domain adaptation is applied to the detector. As a result, the numerical values reported are best viewed as the behaviour of a particular, plausible stereo stack rather than an upper bound on what more specialised algorithms might achieve.
- **Discrete media and lighting conditions.** The media and lighting conditions used are discrete and relatively coarse. Chapter 3 investigates three environments (air, underwater un-

calibrated, underwater recalibrated) and six light colours in air (Ambient, White, Red, Blue, Green, Yellow), with blue light showing catastrophic performance for a smooth clear bottle [33]. Chapter 4 evaluates six environments (Air, Freshwater Day, Freshwater Night, Freshwater+`basicEnhance`, Saltwater, Saltwater+Pellets) [74, 74]. These choices capture key behaviours (e.g., recalibration effects, turbidity and particulate extremes) but do not densely sample intermediate turbidity levels, colour casts, or dynamic lighting (e.g., strong caustics or moving shadows).

- **Narrow enhancement and filtering study.** The enhancement study in Chapter 4 is deliberately minimal: `basicEnhance` applies Canny-based edge blending and a simple depth-aware confidence reweighting with fixed parameters, evaluated only for freshwater recordings [74, 74]. This variant was designed as an illustrative proof-of-concept to demonstrate how the UW-TransStereo benchmark can be used to evaluate inference-time enhancements; its primary goal is to improve detection stability rather than ranging accuracy, so the observed trade-off (improved detection success rates at the cost of increased distance error) is expected by design. No parameter sweeps or comparisons against more sophisticated image enhancement, dehazing, or colour-correction methods are performed. Similarly, temporal smoothing and confidence gating are analysed conceptually and via aggregate metrics, but no full ablation over temporal filters or gating policies is conducted. Conclusions about stability–precision trade-offs are therefore specific to these particular choices.
- **Partial reproducibility for earlier freshwater day sequences.** While UW-TransStereo is released with SVO, spreadsheets, and collated CSVs for all public recordings, some earlier freshwater day sequences used in the analysis were not recorded to `.svo2` and so cannot be replayed at the raw-frame level [74]. Their behaviour is preserved via aggregated tables and plots, but frame-level reproduction of those specific sequences is not possible from the public archive.

6.3 Future work

The limitations above suggest several directions for future research, spanning stereo-informed telepresence, underwater perception, and dataset design.

- **From lab-scale rigs to field deployments.** A natural next step is to translate the stereo-informed MR teleoperation framework and underwater ranging methods from bench-top tanks to more realistic platforms, such as ROVs or AUVs operating in larger pools or open water. This will require handling camera motion, non-planar and cluttered backgrounds, and more variable lighting than considered here, and may motivate tighter integration between stereo perception and vehicle control.

- **Refraction- and geometry-aware underwater stereo.** Across Chapters 3 and 4, the hardest cases are water-filled transparent bottles, which frequently appear see-through to the depth engine and cause the system to lock on to the background [33, 74, 74]. Future work should therefore explore refraction-aware stereo methods that explicitly model layered media and refractive interfaces, and geometry-aware approaches that incorporate object shape priors and multi-view consistency, building on in-air work on transparent-object depth completion and pose estimation [74]. UW-TransStereo provides stereo, depth, and range annotations across multiple media that can serve as a testbed for such methods.
- **Richer media and lighting sweeps.** The saltwater versus freshwater comparison showed that, for this configuration, saltwater can yield significantly better MAE and success rates despite greater turbidity (Section 5.1.2). This suggests that the relationship between medium properties and stereo performance is non-trivial. Future datasets could introduce parametric sweeps over salinity, turbidity, particulate density, and lighting (colour and direction) to better characterise when performance improves or degrades and to support models that estimate or adapt to medium properties jointly with depth.
- **Extended object sets and tasks.** Extending the object set beyond the current bottles to include larger containers, irregular shapes, bags, and other transparent or semi-transparent items would help test the generality of the observed trends and support tasks beyond detection and ranging, such as pose estimation and grasp planning. The capture and logging framework used for UW-TransStereo (paired SVO, spreadsheets, screenshots, and collated CSVs) can be reused for such extensions [74].
- **Algorithmic baselines and multi-modal fusion.** The current analysis focuses on a single detector and depth stack. Future work should broaden algorithmic baselines (different detectors, depth settings, and fusion strategies) to understand which components contribute most to robustness in different media. Integrating additional modalities such as time-of-flight, structured light, or polarisation imaging with stereo could improve performance in high-turbidity or strongly refractive scenarios. Temporal fusion and tracking, beyond the

conceptual discussion in this thesis, also merit systematic evaluation.

- **Closing the loop to telepresence.** Finally, there is an opportunity to close the loop between the MR telepresence work in Chapter 3 and the dataset-driven analyses in Chapter 4. Methods validated on UW-TransStereo could be integrated back into the MR teleoperation pipeline, using improved depth and detection under challenging media to drive more reliable distance cues, sonification, and manipulation support in real time.

In summary, this thesis demonstrates that stereo vision can provide useful distance cues for augmented telepresence and that carefully designed underwater datasets such as UW-TransStereo can reveal, and help quantify, the limits of current methods for transparent-object detection and ranging. At the same time, the results highlight that robust underwater stereo for refractive objects remains an open challenge, and that future progress will depend on combining improved physical modelling, richer datasets, and closer integration between perception and control.

Appendix A

Error Distribution Visualisations

This appendix presents detailed visualisations of error distributions that complement the summary statistics in Section 5.1.3. The figures below provide diagnostic insight into the shape of ranging errors across environments and bottle types, revealing patterns such as bimodality (indicating distinct failure modes) and heavy tails (indicating outlier-prone distributions).

A.1 Q-Q Plots for Normality Assessment

Figure A.1 presents quantile-quantile (Q-Q) plots comparing observed error quantiles against theoretical normal quantiles for each environment, aggregated across bottle types. Deviations from the diagonal reference line indicate departure from normality: S-shaped curves suggest heavy tails (more extreme values than expected under normality), while systematic departures above or below the line indicate skewness.

A.2 Violin Plots of Error Distributions

Figures A.2 and A.3 present violin plots showing the full error distribution for each environment–bottle type combination. The width of each violin indicates the density of errors at that value; wider sections correspond to more frequent errors. The horizontal red line at zero represents perfect accuracy (no bias).

These plots reveal distributional features not captured by summary statistics:

- **Bimodality:** The filled smooth bottle in underwater environments exhibits two distinct

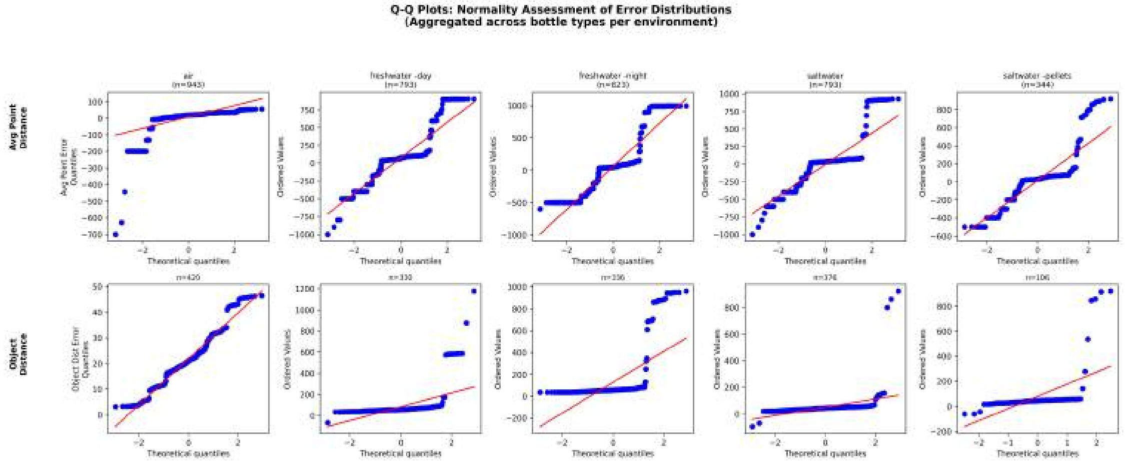


Figure A.1: Q-Q plots comparing observed error quantiles against theoretical normal quantiles for each environment (columns) and range estimator (rows: Avg Point Distance, Object Distance). The diagonal line represents perfect normality; deviations indicate non-Gaussian error distributions. S-shaped curves (visible in most panels) indicate heavy tails characteristic of stereo matching failures. The filled smooth bottle contributes most heavily to these deviations due to rear-wall locking.

modes—one near zero (successful surface detection) and one at large positive values (rear-wall locking). This directly visualises the index-matching failure mode.

- **Asymmetry:** Most distributions are asymmetric, with longer tails extending toward positive errors (overestimation), consistent with the systematic rear-wall bias discussed in Section 5.1.1.
- **Environmental gradient:** Error distributions progressively widen from Air (tightest) through Freshwater to Saltwater with Pellets (widest), reflecting the increasing challenge of stereo matching under turbidity and backscatter.

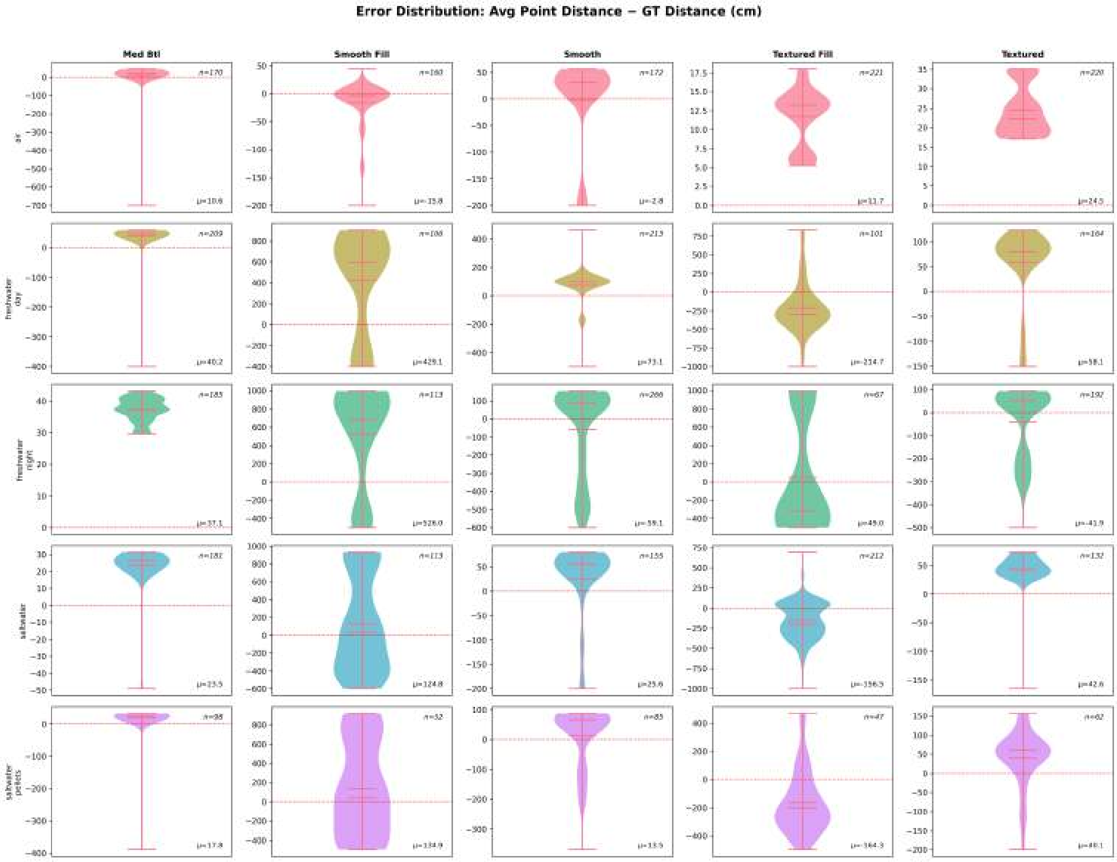


Figure A.2: Violin plots of Avg Point Distance error (predicted minus ground truth, in cm) by environment (rows) and bottle type (columns). Each violin shows the full error distribution; width indicates density. The horizontal line at zero represents ideal (unbiased) estimation. The filled smooth bottle exhibits bimodal distributions in underwater environments, with errors clustering near zero (successful surface detection) and at large positive values (rear-wall locking). Sample sizes (n) and mean bias (μ) are annotated per cell.

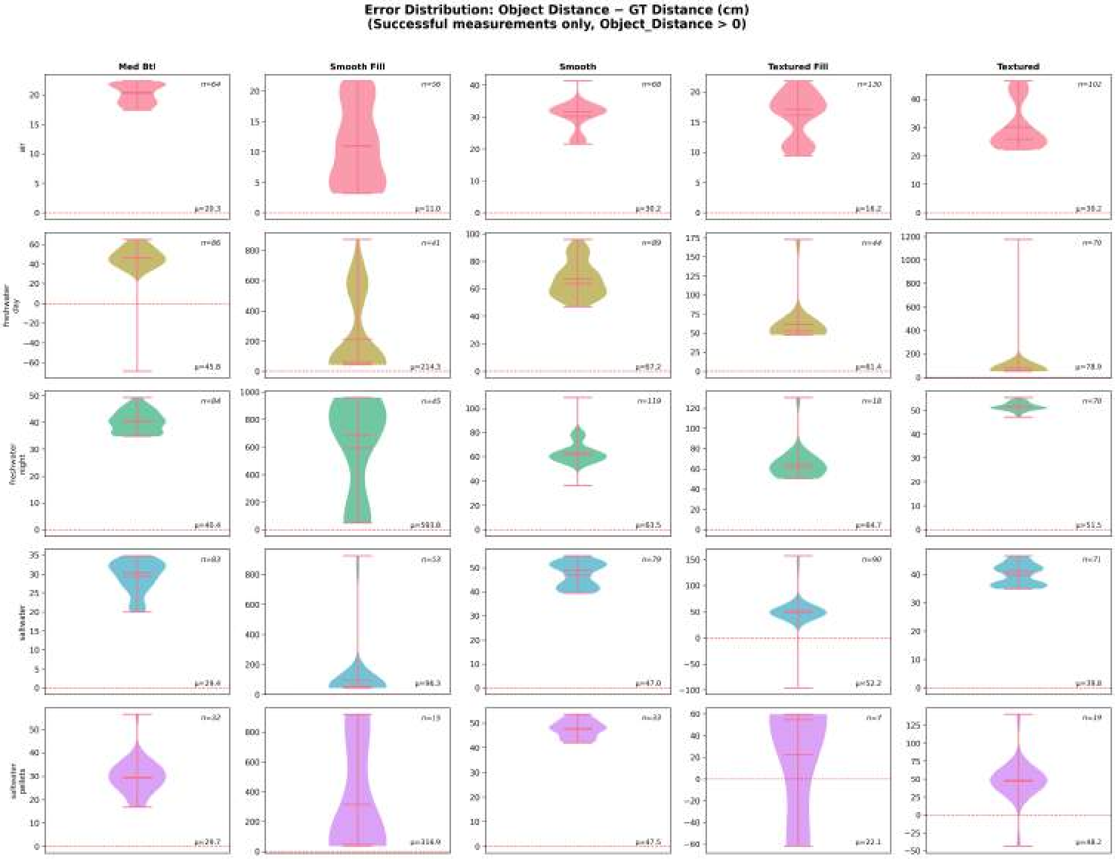


Figure A.3: Violin plots of Object Distance error by environment and bottle type (successful measurements only, Object Distance > 0). Object Distance errors show consistently positive bias across all conditions, indicating systematic overestimation. The distributions are generally narrower than Avg Point Distance due to the SDK's spatial filtering, but bimodal patterns persist for the filled smooth bottle in underwater environments.

Appendix B

Object Dimension Analysis

This appendix presents supplementary analysis of the ZED SDK's 3D object dimension estimates (width, height, depth) compared against ground-truth measurements for the five bottle types across all test environments. The dimension data was logged alongside distance measurements (Section 4.5.1) but represents a distinct capability: estimating object *size* rather than object *range*.

B.1 Methodology

Object dimensions are computed by the ZED SDK's bounding-box estimator, which projects the 2D detection region into 3D space using the stereo depth map and returns width, height, and depth in millimetres. Ground-truth dimensions were measured manually for each bottle type (Table 4.3 in Section 4.4.2). A measurement is considered *successful* if all three dimensions return non-zero values; measurements where any dimension is zero (indicating SDK failure) are excluded from accuracy analysis.

Outlier filtering. Preliminary analysis revealed that approximately 11% of successful measurements contain gross errors where one or more dimensions deviate by more than 2× from ground truth (e.g., width estimates of 2208 mm for a 50 mm bottle). These outliers arise from stereo matching failures in refractive or low-contrast regions and would dominate summary statistics if included. Following standard practice for robust error analysis, we report metrics under two regimes:

- **Raw metrics:** All successful measurements (non-zero dimensions), including outliers.
- **Filtered metrics:** Measurements where all dimension ratios (measured/GT) fall within [0.5, 2.0], excluding gross errors.

B.2 Summary Results

Figure B.1 presents a comprehensive 8-panel summary of dimension accuracy. Key findings:

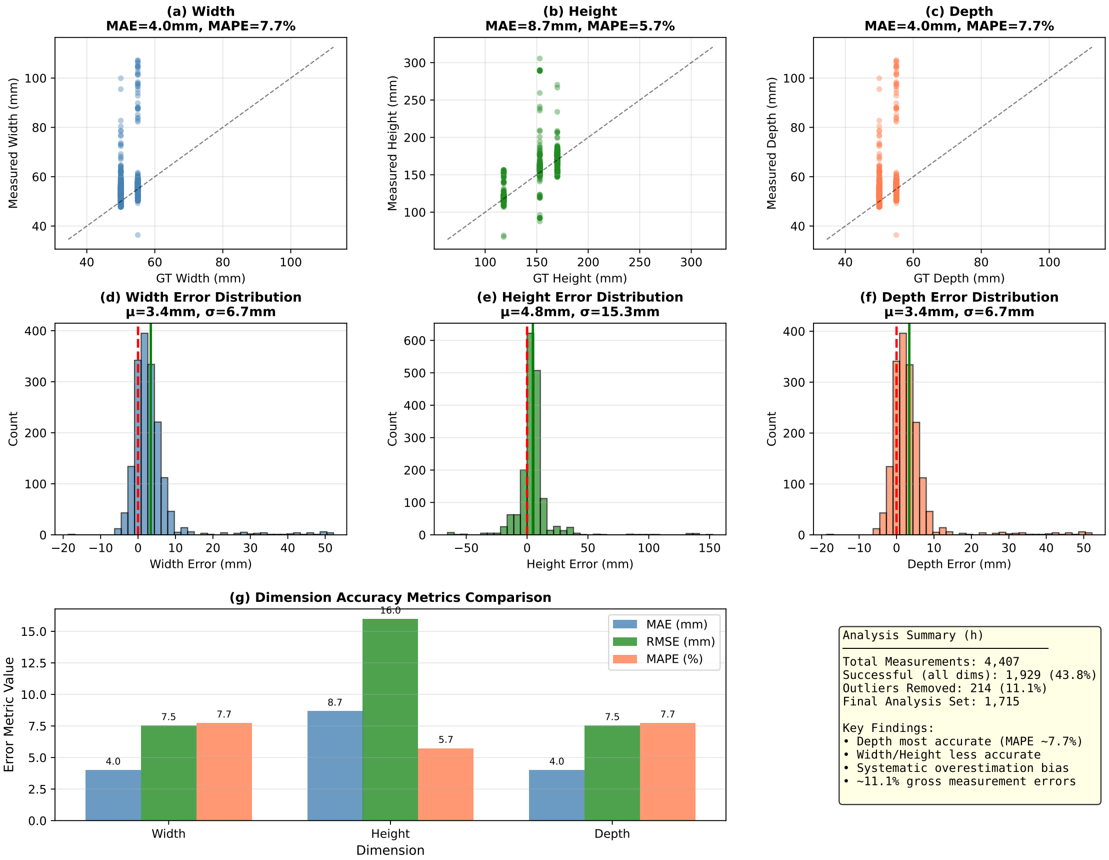
- **Success rate:** 43.8% of detections return valid 3D dimensions—similar to the Object Distance success rate (Table 5.1), indicating that dimension estimation faces the same underlying challenges as ranging.
- **Filtered accuracy:** After removing gross outliers, all three dimensions achieve 5–8% Mean Absolute Percentage Error (MAPE): Width 7.7%, Height 5.7%, Depth 7.7%. This demonstrates that when the stereo pipeline succeeds, dimension estimates are reasonably accurate.
- **Gross outlier rate:** 11.1% of successful measurements (214 of 1,929) exceed the 2× threshold, representing cases where the SDK produces wildly incorrect dimension estimates—an important caveat for applications relying on 3D object sizing.
- **Systematic bias:** All dimensions exhibit positive bias (overestimation): Width +3.4 mm, Height +4.8 mm, Depth +3.4 mm (filtered). This mirrors the positive bias observed in distance measurements (Section 5.1.1).
- **Environmental gradient:** Dimension accuracy degrades from Air through Freshwater to Saltwater with Pellets, following the same pattern as distance measurements. The freshwater-basicEnhance condition shows the highest filtered success rate (51.5%), while saltwater-pellets shows the lowest (33.1%).

B.3 Detailed Metrics

Table B.1 compares raw and filtered metrics for each dimension, quantifying the impact of gross outliers on summary statistics.

B.4 Implications

The dimension analysis complements the ranging results in Chapter 5 by characterising the ZED SDK's full 3D pose estimation capability for underwater transparent objects. The key implica-



```
Analysis Summary (h)
─────────────────────────────
Total Measurements: 4,407
Successful (all dims): 1,929 (43.8%)
Outliers Removed: 214 (11.1%)
Final Analysis Set: 1,715

Key Findings:
• Depth most accurate (MAPE ~7.7%)
• Width/Height less accurate
• Systematic overestimation bias
• ~11.1% gross measurement errors
```

Figure B.1: Comprehensive dimension analysis summary. Panels (a–c): Scatter plots of measured vs ground-truth dimensions for Width, Height, and Depth (outliers removed), with MAE and MAPE annotations. Panels (d–f): Error distribution histograms showing mean (μ, green) and zero reference (red). Panel (g): Bar chart comparing MAE, RMSE, and MAPE across dimensions. Panel (h): Summary statistics including total measurements, outlier rate, and key findings. All metrics computed on filtered data (outliers removed) unless otherwise noted. $n = 1{,}715$ valid measurements after filtering.

Table B.1: Dimension measurement metrics: raw (all successful) vs filtered (outliers removed). Outlier threshold: measured/GT ratio outside [0.5, 2.0]. The dramatic difference between raw and filtered MAPE for Width (181% vs 7.7%) illustrates the impact of gross measurement errors on summary statistics.

Dimension	n (raw)	Outliers	Raw MAPE	Filtered MAPE	Filtered MAE	Filtered Bias
Width	1,929	214 (11.1%)	181.1%	7.7%	4.0 mm	+3.4 mm
Height	1,929	214 (11.1%)	45.0%	5.7%	8.7 mm	+4.8 mm
Depth	1,929	214 (11.1%)	19.6%	7.7%	4.0 mm	+3.4 mm

tions are:

1. **Comparable success rates:** Dimension estimation and ranging share similar success rates (44%), suggesting they face common underlying challenges (stereo correspondence failures, refraction, low contrast).
2. **Reasonable filtered accuracy:** When successful and non-outlier, dimension estimates are accurate within 5–8% MAPE—sufficient for coarse object sizing or classification by size category, but potentially insufficient for precision manipulation.
3. **Gross error caveat:** The 11% outlier rate represents a significant failure mode where dimensions are wildly incorrect. Applications using dimension data should implement outlier detection (e.g., temporal consistency checks, geometric constraints) to filter unreliable estimates.
4. **Depth most reliable:** Among the three dimensions, Depth (aligned with the camera optical axis) shows the lowest raw MAPE (19.6%), suggesting that depth-aligned measurements are more robust to stereo matching errors than perpendicular (Width/Height) measurements.

B.5 Supplementary Visualisations

This section presents additional visualisations that complement the filtered RMSE heatmap in Figure 5.22.

B.5.1 Raw RMSE Heatmap (Including Outliers)

Figure B.2 shows dimension RMSE computed over *all* successful measurements, including gross outliers. The dramatically higher values (100–1400 mm vs 1–25 mm in the filtered version) illustrate the impact of the 11% outlier rate on summary statistics. Width shows the largest raw RMSE because it is most susceptible to outliers where the SDK returns bounding-box dimensions spanning the entire detection region rather than the object itself.

B.5.2 Success Rate Heatmap

Figure B.3 shows the proportion of detections returning valid (non-zero) 3D dimensions for each environment and bottle type. Success rates range from 15–60%, with the environmental gradient matching distance findings: Air and freshwater-basicEnhance achieve the highest rates, while

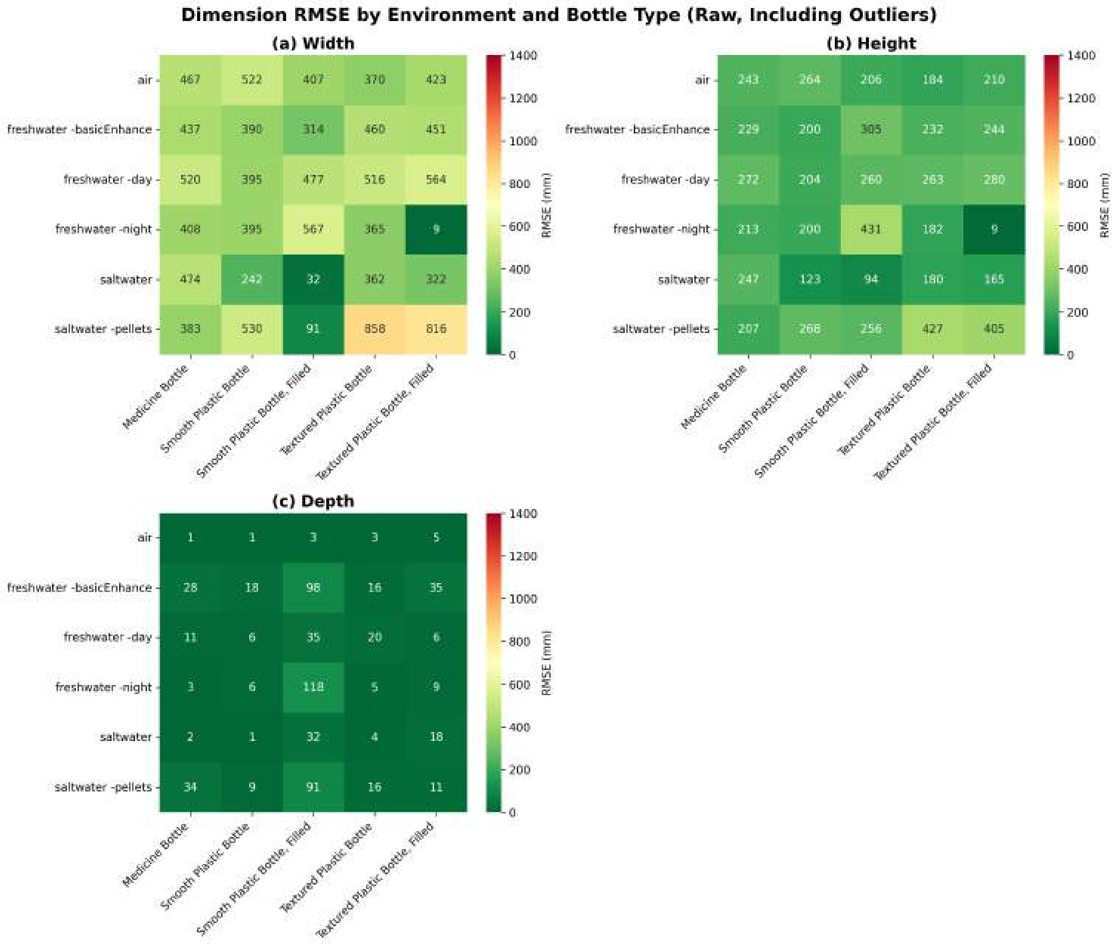


Figure B.2: Dimension RMSE by environment and bottle type (raw, including outliers). Compare with Figure 5.22 (outliers removed). The order-of-magnitude difference illustrates the impact of gross measurement errors: Width raw RMSE reaches 600–1400 mm, while filtered RMSE is 1–25 mm. These outliers arise from stereo matching failures where the SDK reports wildly incorrect object dimensions.

saltwater-pellets shows the lowest. The filled textured bottle in saltwater-pellets exhibits particularly low success rates (15–21%), indicating that suspended particulates severely degrade the SDK's ability to compute 3D object extents.

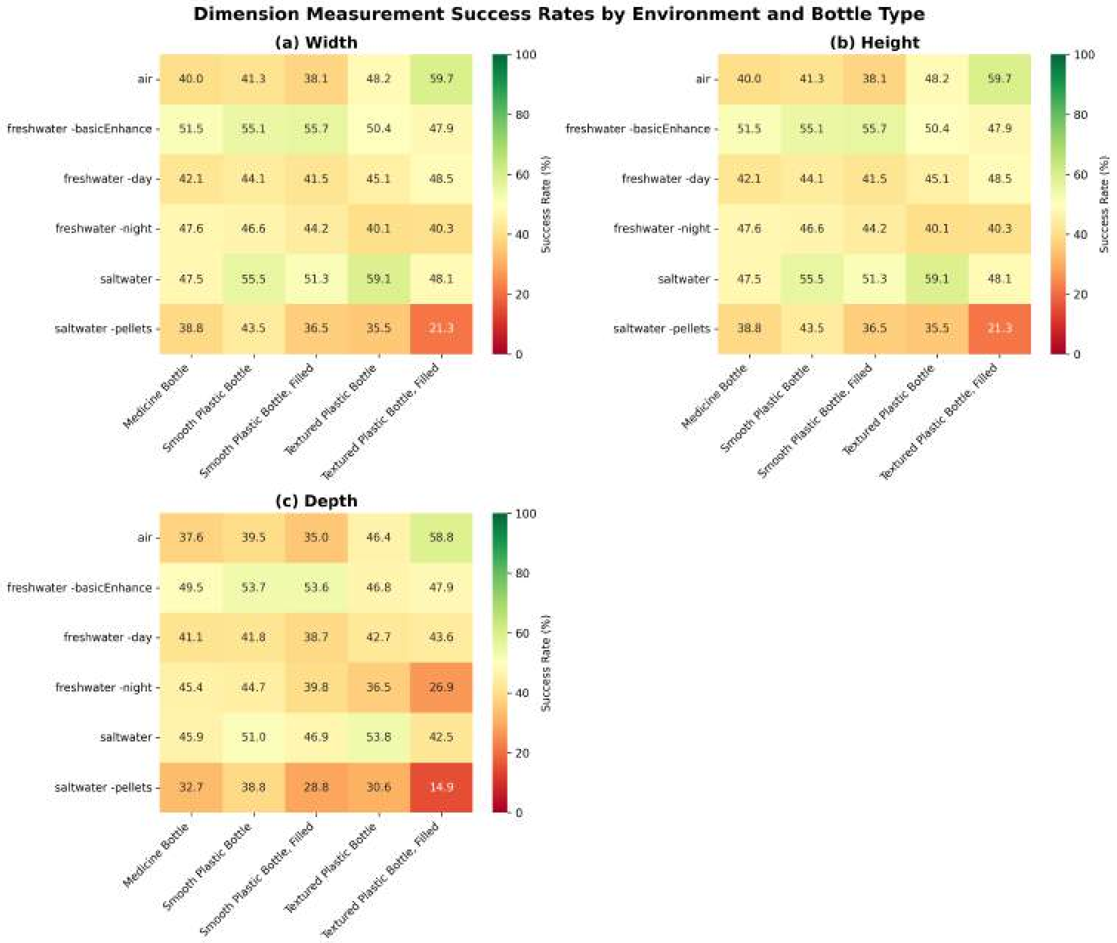


Figure B.3: Dimension measurement success rates by environment and bottle type. Success rate is defined as the proportion of detections returning non-zero values for all three dimensions. The overall success rate (43.8%) closely matches Object Distance success rates (Table 5.1), indicating that dimension estimation and ranging face common underlying challenges. Saltwater-pellets shows the lowest rates across all bottle types.

Bibliography

[1] Elmehdi Adil, Mohammed Mikou, and Ahmed Mouhsen. A novel algorithm for distance measurement using stereo camera. *CAAI Transactions on Intelligence Technology*, 7(2):177–186, 2022.

[2] Amit Agrawal, Srikumar Ramalingam, Yuichi Taguchi, and Visesh Chari. A theory of multi-layer flat refractive geometry. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3346–3353.

[3] Derya Akkaynak and Tali Treibitz. A Revised Underwater Image Formation Model. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6723–6732.

[4] Derya Akkaynak and Tali Treibitz. Sea-Thru: A Method for Removing Water From Underwater Images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1682–1691.

[5] J. Atherton and Michael Goodrich. Supporting remote and mobile manipulation with an ecological augmented virtuality interface.

[6] A. Bettini, P. Marayong, S. Lang, A.M. Okamura, and G.D. Hager. Vision-assisted control for manipulation using virtual fixtures. 20(6):953–966.

[7] Gideon Billings, Richard Camilli, and Matthew Johnson-Roberson. Hybrid Visual SLAM for Underwater Vehicle Manipulator Systems.

[8] Josep Bosch, Klemen Istenio, Nuno Gracias, Rafael Garcia, and Pere Ridao. Omnidirectional Multicamera Video Stitching Using Depth Maps. 45(4):1337–1352.

[9] F. Bruno, G. Bianco, M. Muzzupappa, S. Barone, and A. V. Razionale. Experimentation of structured light and stereo vision for underwater 3D reconstruction. 66(4):508–518.

[10] R. Budwig. Refractive index matching methods for liquid flow investigations. *Experiments in Fluids*, 17(5):350–355, 1994.

[11] Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. ClearPose: Large-scale Transparent Object Dataset and Benchmark.

[12] Xiaoxue Chen, Junchen Liu, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. NeRRF: 3D Reconstruction and View Synthesis for Transparent and Specular Objects with Neural Refractive-Reflective Fields.

[13] Copernicus Marine Service. From plastic... to marine pollution. `https://marine.copernicus.eu/explainers/phenomena-threats/plastic-pollution/from-plastic-marine-pollution`, 2024. Accessed: 2024.

[14] Xiwen Deng, Tao Liu, Shuangyan He, Xinyao Xiao, Peiliang Li, and Yanzhen Gu. An underwater image enhancement model for domain adaptation. 10.

[15] Mica Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 32.

[16] Marcus Eriksen, Laurent C. M. Lebreton, Henry S. Carson, Martin Thiel, Charles J. Moore, Jose C. Borerro, François Galgani, Peter G. Ryan, and Julia Reisser. Plastic pollution in the world's oceans: More than 5 trillion plastic pieces weighing over 250,000 tons afloat at sea. *PLOS ONE*, 9(12):e111913, 2014.

[17] Michael Fulton, Jungseok Hong, Md Jahidul Islam, and Junaed Sattar. Robotic Detection of Marine Litter Using Deep Visual Detection Models. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5752–5758.

[18] Michael Fulton, Jungseok Hong, Md Jahidul Islam, and Junaed Sattar. Robotic detection of marine litter using deep visual detection models. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE.

[19] Michael S. Fulton, Jungseok Hong, and Junaed Sattar. Trash-ICRA19: A Bounding Box Labeled Dataset of Underwater Trash.

[20] Arturo Gomez Chavez, Andrea Ranieri, Davide Chiarella, Enrica Zereik, Anja Babić, and Andreas Birk. CADDY Underwater Stereo-Vision Dataset for Human–Robot Interaction (HRI) in the Context of Diver Activities. 7(1):16.

[21] Salma P. González-Sabbagh and Antonio Robles-Kelly. A Survey on Underwater Computer Vision. page 3578516.

[22] Honey Gupta and Kaushik Mitra. Unsupervised Single Image Underwater Depth Estimation.

[23] Oksana Hagen, Amir Aly, Ray Jones, Marius Varga, and Dena Bazazian. Beyond the surface: A scoping review of vision-based underwater experience technologies and user studies. 2(1):1–24.

[24] Butler Hine, Carol Stoker, Michael Sims, Daryl Rasmussen, Terrence Fong, Jay Steele, and Don Barch. The application of telepresence and virtual reality to subsea exploration. *Proc. ROV '94*, 1994.

[25] Chao Hu, Shiqiang Zhu, and Wei Song. Real-time Underwater 3D Reconstruction Based on Monocular Image. In *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1239–1244.

[26] Xia Hua, Xiaopeng Cui, Xinghua Xu, Shaohua Qiu, Yingjie Liang, Xianqiang Bao, and Zhong Li. Underwater object detection algorithm based on feature enhancement and progressive dynamic aggregation strategy. 139:109511.

[27] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

[28] Dong Huo, Jian Wang, Yiming Qian, and Yee-Hong Yang. Glass Segmentation with RGB-Thermal Image Pairs. 32:1911–1926.

[29] Guanying Huo, Ziyin Wu, Jiabiao Li, and Shoujun Li. Underwater Target Detection and 3D Reconstruction System Based on Binocular Vision. 18(10):3570.

[30] Guanying Huo, Ziyin Wu, Jiabiao Li, and Shoujun Li. Underwater target detection and 3d reconstruction system based on binocular vision. *Sensors*, 18(10):3570, 2018.

[31] Ivo Ihrke, Kiriakos N. Kutulakos, Hendrik P. A. Lensch, Marcus Magnor, and Wolfgang Heidrich. Transparent and specular object reconstruction. In *Eurographics 2007, State of the Art Reports*, 2007.

[32] Md Jahidul Islam, Jungseok Hong, and Junaed Sattar. Person Following by Autonomous Robots: A Categorical Overview.

[33] Adrian Kaehler and Gary Bradski. *Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library*. O'Reilly Media, Inc., 2016.

[34] Kinga Korniejenko and Bartosz Kontny. The Usage of Virtual and Augmented Reality in Underwater Archeology. 14(18):8188.

[35] Clayton Kunz and Hanumant Singh. Hemispherical refraction and camera calibration in underwater vision. In *OCEANS 2008*, pages 1–7.

[36] D.M. Lane, J.B.C. Davies, G. Casalino, G. Bartolini, G. Cannata, G. Veruggio, M. Canals, C. Smith, D.J. O'Brien, M. Pickett, G. Robinson, D. Jones, E. Scott, A. Ferrara, D. Angelleti, M. Coccoli, R. Bono, P. Virgili, R. Pallas, and E. Gracia. Amadeus: advanced manipulation for deep underwater sampling. *IEEE Robotics & Automation Magazine*, 4(4):34–45, 1997. OA status: green$_{p}ublished$.

[37] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep Learning for Detecting Robotic Grasps.

[38] Deborah Levy, Amit Peleg, Naama Pearl, Dan Rosenbaum, Derya Akkaynak, Simon Korman, and Tali Treibitz. SeaThru-NeRF: Neural Radiance Fields in Scattering Media.

[39] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. 98:107038.

[40] Jie Li, Katherine A. Skinner, Ryan M. Eustice, and Matthew Johnson-Roberson. WaterGAN: Unsupervised Generative Network to Enable Real-time Color Correction of Monocular Underwater Images. pages 1–1.

[41] Na Li, Ziqiang Zheng, Shaoyong Zhang, Zhibin Yu, Haiyong Zheng, and Bing Zheng. The Synthesis of Unpaired Underwater Images Using a Multistyle Generative Adversarial Network. 6:54241–54257.

[42] Jiaying Lin, Zebang He, and Rynson W. H. Lau. Rich Context Aggregation With Reflection Prior for Glass Surface Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[44] Jeffrey I. Lipton, Aidan J. Fay, and Daniela Rus. Baxter's homunculus: Virtual reality spaces for teleoperation in manufacturing. *IEEE Robotics and Automation Letters*, 3(1):179–186, 2018.

[45] Chongwei Liu, Haojie Li, Shuchang Wang, Ming Zhu, Dong Wang, Xin Fan, and Zhihui Wang. A Dataset and Benchmark of Underwater Object Detection for Robot Picking. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6.

[46] Xingyu Liu, Shun Iwase, and Kris M. Kitani. StereOBJ-1M: Large-scale Stereo Image Dataset for 6D Object Pose Estimation.

[47] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. KeyPose: Multi-View 3D Labeling and Keypoint Estimation for Transparent Objects.

[48] Salvatore Livatino, Dario C. Guastella, Giovanni Muscato, Vincenzo Rinaldi, Luciano Cantelli, Carmelo D. Melita, Alessandro Caniglia, Riccardo Mazza, and Gianluca Padula. Intuitive robot teleoperation through multi-sensor informed mixed reality visual aids. *IEEE Access*, 9:25795–25808, 2021.

[49] Dario Lodi Rizzini, Fabjan Kallasi, Jacopo Aleotti, Fabio Oleari, and Stefano Caselli. Integration of a stereo vision system into an autonomous underwater vehicle for pipe manipulation tasks. 58:560–571.

[50] Dario Lodi Rizzini, Fabjan Kallasi, Jacopo Aleotti, Fabio Oleari, and Stefano Caselli. Integration of a stereo vision system into an autonomous underwater vehicle for pipe manipulation tasks. *Computers & Electrical Engineering*, 58:560–571, 2017.

[51] Huimin Lu, Yujie Li, Tomoki Uemura, Hyoungseop Kim, and Seiichi Serikawa. Low illumination underwater light field images reconstruction using deep convolutional neural networks. 82:142–148.

[52] Cai Luo, Jihua Wu, Shixin Sun, and Peng Ren. TransCODNet: Underwater Transparently Camouflaged Object Detection via RGB and Event Frames Collaboration. 9(2):1444–1451.

[53] Yunpeng Ma, Yi Wu, Qingwu Li, Yaqin Zhou, and Dabing Yu. ROV-based binocular vision system for underwater structure crack detection and width measurement. 82(14):20899–20923.

[54] Giacomo Marani, Song K. Choi, and Junku Yuh. Underwater autonomous manipulation for intervention missions auvs. *Ocean Engineering*, 36(1):15–23, 2009.

[55] Inc Matterport. Mask r-cnn for object detection and segmentation, 2023-09-25T14:48:29Z 2023.

[56] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass Segmentation using Intensity and Spectral Polarization Cues. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12612–12621.

[57] Mehdi Mousavi, Shardul Vaidya, Razat Sutradhar, and Ashwin Ashok. OpenWaters: Photorealistic Simulations For Underwater Computer Vision. In *The 15th International Conference on Underwater Networks & Systems*, pages 1–5. ACM.

[58] Fickrie Muhammad, Rifakhryza A. Mugiaraya, Gabriella Alodia, and Harald Sternberg. A Comparative Analysis of Refraction-Aware SfM, Hierarchical Localization, and Gaussian Splatting for Underwater 3D Reconstruction.

[59] Karen Panetta, Chen Gao, and Sos Agaian. Human-Visual-System-Inspired Underwater Image Quality Measures. 41(3):541–551.

[60] Mostafa Parchami and Gian-Luca Mariottini. A comparative study on 3-d stereo reconstruction from endoscopic images. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM.

[61] Matteo Poggi and Stefano Mattoccia. Learning to Predict Stereo Reliability Enforcing Local Consistency of Confidence Maps. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4541–4550.

[62] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Quantitative Evaluation of Confidence Measures in a Machine Learning World. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5238–5247.

[63] Dimitris V. Politikos, Elias Fakiris, Athanasios Davvetas, Iraklis A. Klampanos, and George Papatheodorou. Automatic detection of seafloor marine litter using towed camera images and deep learning. 164:111974.

[64] Dimitris V. Politikos, Elias Fakiris, Athanasios Davvetas, Iraklis A. Klampanos, and George Papatheodorou. Automatic detection of seafloor marine litter using towed camera images and deep learning. *Marine Pollution Bulletin*, 164:111974, 2021.

[65] R. L. Pyle, R. Boland, H. Bolick, B. W. Bowen, C. J. Bradley, C. Kane, R. K. Kosaki, R. Langston, K. Longenecker, A. Montgomery, F. A. Parrish, B. N. Popp, J. Rooney, C. M. Smith, D. Wagner, and H. L. Spalding. A comprehensive investigation of mesophotic coral ecosystems in the hawaiian archipelago. *PeerJ*, 4:e2475, 2016. Pyle, Richard L Boland, Raymond Bolick, Holly Bowen, Brian W Bradley, Christina J Kane, Corinne Kosaki, Randall K Langston, Ross Longenecker, Ken Montgomery, Anthony Parrish, Frank A Popp, Brian N Rooney, John Smith, Celia M Wagner, Daniel Spalding, Heather L eng 2016/10/21 PeerJ. 2016 Oct 4;4:e2475. doi: 10.7717/peerj.2475. eCollection 2016.

[66] Juan Roldán, Elena Peña-Tapia, Andrés Martín-Barrio, Miguel Olivares-Méndez, Jaime Del Cerro, and Antonio Barrientos. Multi-robot interfaces and operator situational awareness:

Study of the impact of immersion and prediction. *Sensors*, 17(8):1720, 2017. OA status: green$_p$*ublished*.

[67] Shreeyak S. Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. ClearGrasp: 3D Shape Estimation of Transparent Objects for Manipulation.

[68] Luigi Scarfone, Rosario Aiello, Umberto Severino, Loris Barbieri, and Fabio Bruno. Online 3D Reconstruction in Underwater Environment Using a Low-Cost Depth Camera. In Caterina Rizzi, Francesca Campana, Michele Bici, Francesco Gherardini, Tommaso Ingrassia, and Paolo Cicconi, editors, *Design Tools and Methods in Industrial Engineering II*, pages 237–244. Springer International Publishing.

[69] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/3):7–42, 2002.

[70] Anne Sedlazeck and Reinhard Koch. Perspective and Non-perspective Camera Models in Underwater Imaging – Overview and Error Analysis.

[71] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8):11–11, 2011.

[72] Hyungwon Shim, Bong-Huan Jun, Pan-Mook Lee, Hyuk Baek, and Jihong Lee. Workspace control system of underwater tele-operated manipulators on an rov. *Ocean Engineering*, 37(11):1036–1047, 2010.

[73] Mark Shortis. Calibration Techniques for Accurate Measurements by Underwater Camera Systems. 15(12):30810–30826.

[74] Aaron L. Smiles, Changjae Oh, and Ildar Farkhatdinov. Multi-environment stereo dataset for transparent object detection & ranging (uw-transstereo).

[75] Aaron L. Smiles, Changjae Oh, and Ildar Farkhatdinov. A preliminary study on underwater transparent objects detection with stereo vision: Air vs freshwater. In *Proc. 13th Int. Conf. Robot Intelligence Technology and Applications (RiTA)*.

[76] Wentao Sun, Yiming Bi, Rohan Mukherjee, Patrick McCarthy, Hisashi Ishida, and Shuran Song. TaTa: Visual-tactile fusion for transparent object grasping in clutter. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[77] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer International Publishing, Cham, 2022.

[78] Camilo Sánchez-Ferreira, Jones Y. Mori, Mylène C. Q. Farias, and Carlos H. Llanos. A real-time stereo vision system for distance measurement and underwater image restoration. 38(7):2039–2049.

[79] Camilo Sánchez-Ferreira, Jones Y. Mori, Mylène C. Q. Farias, and Carlos H. Llanos. A real-time stereo vision system for distance measurement and underwater image restoration. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 38(7):2039–2049, 2016.

[80] Hakon Teigland, Vahid Hassani, and Ments Tore Moller. Operator focused automation of rov operations. In *2020 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV)*. IEEE.

[81] Tali Treibitz, Yoav Schechner, Clayton Kunz, and Hanumant Singh. Flat Refractive Geometry. 34(1):51–65.

[82] Peter Tueller, Raghav Maddukuri, Patrick Paxson, Vivaswat Suresh, Arjun Ashok, Madison Bland, Ronan Wallace, Julia Guerrero, Brice Semmens, and Ryan Kastner. FishSense: Underwater RGBD Imaging for Fish Measurement. In *OCEANS 2021: San Diego – Porto*, pages 1–5.

[83] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields.

[84] Cong Wang, Qifeng Zhang, Sen Lin, Wentao Li, Xiaohui Wang, Yunfei Bai, and Qiyan Tian. Research and experiment of an underwater stereo vision system. In *OCEANS 2019 - Marseille*, pages 1–5, Marseille, France. IEEE.

[85] Cong Wang, Qifeng Zhang, Sen Lin, Wentao Li, Xiaohui Wang, Yunfei Bai, and Qiyan Tian. Research and Experiment of an Underwater Stereo Vision System. In *OCEANS 2019 - Marseille*, pages 1–5. IEEE.

[86] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16606–16616.

[87] Junjie Wen, Jinqiang Cui, Zhenjun Zhao, Ruixin Yan, Zhi Gao, Lihua Dou, and Ben M. Chen. SyreaNet: A Physically Guided Underwater Image Enhancement Framework Integrating Synthetic and Real Images. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5177–5183.

[88] Lucy C. Woodall, Anna Sanchez-Vidal, Miquel Canals, Gordon L. J. Paterson, Rachel Coppock, Victoria Sleight, Antonio Calafat, Alex D. Rogers, Bhavani E. Narayanaswamy, and Richard C. Thompson. The deep sea is a major sink for microplastic debris. *Royal Society Open Science*, 1(4):140317, 2014.

[89] Yanli Wu, Rui Nian, and Bo He. 3D reconstruction model of underwater environment in stereo vision system. In *2013 OCEANS - San Diego*, pages 1–4.

[90] Hu X and Mordohai P. A quantitative evaluation of confidence measures for stereo vision.

[91] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting Transparent Objects in the Wild.

[92] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting Transparent Object in the Wild with Transformer.

[93] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. DepthSplat: Connecting Gaussian Splatting and Depth.

[94] Daniel Yang, John J. Leonard, and Yogesh Girdhar. SeaSplat: Representing Underwater Scenes with 3D Gaussian Splatting and a Physically Grounded Image Formation Model.

[95] Miao Yang and Arcot Sowmya. An Underwater Color Image Quality Evaluation Metric. 24(12):6062–6071.

[96] Tian Ye, Sixiang Chen, Yun Liu, Yi Ye, Erkang Chen, and Yuche Li. Underwater Light Field Retention : Neural Rendering for Underwater Imaging. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 487–496. IEEE.

[97] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. Trans4Trans: Efficient Transformer for Transparent Object and Semantic Scene Segmentation in Real-World Navigation Assistance. 23(10):19173–19186.

[98] Zheming Zhou, Xiaotong Chen, and Odest Chadwicke Jenkins. LIT: Light-Field Inference of Transparency for Refractive Object Localization. 5(3):4548–4555.

[99] Antun uraš, Ben J. Wolf, Athina Ilioudi, Ivana Palunko, and Bart De Schutter. A Dataset for Detection and Segmentation of Underwater Marine Debris in Shallow Waters. 11(1):921.