

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



Underwater Tri-dimensional Scene Understanding for an Eye-in-Hand Perception System

Pedro Nuno Barbosa Leite

DOCTORAL PROGRAM IN ELECTRICAL AND COMPUTERS ENGINEERING

Supervisor: Andry Maykol Gomes Pinto

May 13, 2025

Abstract

Offshore wind energy has emerged as a leading renewable technology due to higher capacity factors and larger turbine sizes compared to onshore installations, requiring effective maintenance procedures to operate in harsh maritime environments. This leads to cyclical inspection needs that assess structural integrity and detect potential defects in submerged components. While robotic solutions offer safer alternatives to manual operations that expose divers and support personnel to significant risks, the perceptual capabilities integrated in these systems provide degraded uni-modality representations, limited by data acquisition range and occlusions. Current underwater imaging sensors lack mitigation strategies to reduce the impact of light absorption, scattering, and suspended particles. These not only reduce visibility, limit contrast, and introduce haze, but also critically hinder the acquisition of high-fidelity tri-dimensional information crucial for autonomous structural assessment.

This research focuses on advancing underwater imaging solutions tailored for eye-in-hand perception. The MARESy hybrid imaging sensor is developed to meet close-range inspection requirements, gathering high-fidelity information through both passive stereo imaging and active laser triangulation. A data-driven range-gated acquisition scheme mitigates the impact of sub-sea phenomena, while information fusion principles are introduced to leverage the complementary strengths of each imaging modality. This approach enables the integration of textured data, tri-dimensional point clouds, and motion perception principles into unified representations of the underwater scene.

MARESy demonstrates its effectiveness through embedded processing and millimeter-precise measurements within its optimal operating range, maintaining reliable performance at the ATLANTIS Coastal Testbed despite challenges including biofouling, varying light conditions, and heavy marine snow. Vision-based tri-dimensional information gathered underwater exhibits distinct error profiles and data characteristics: passive stereo provides dense and textured information but its metric performance degrades with distance, while active laser triangulation estimates a sparse set of precise measurements with limited spatial coverage. Novel fusion methodologies are introduced to exploit these complementary characteristics, representing the first approaches to integrate heterogeneous 3D data from underwater imaging techniques. The Joint Masked Regression (JMR) algorithm propagates laser-based measurements through regression curves across homogeneous regions, while the underWater HEterogenous 3d information fusion Architecture (RHEA) jointly extracts features from coupled point cloud inputs through a synthetic-to-real training approach. Both methodologies achieve significant improvements in controlled trials, with accuracy gains of 34.9% and 52.5% over baseline stereo reconstructions, maintaining performance in harsh near-real conditions at the ATLANTIS Test Center. Stereo-based depth estimation remains fundamentally constrained by feature matching ambiguities and continuous accuracy degradation. This research addresses these limitations through multimodal data fusion principles, introducing fusiNg tEXture with apparent motion information for underwater Scene reCOnstruction (NESO) architecture to obtain dense textured representations through the fusion of image data with appar-

ent motion perception principles. The network leverages an attention-based pyramidal design that preserves high-level features by reintroducing them throughout multiple decoding stages. Validation with sacrificial anode inspections demonstrates absolute errors down to 0.0016 m, although predictions require refinement outside the training metric range. Integration with sparse laser data addresses this limitation, achieving metric improvements up to 0.1124 m while preserving relative depth relationships.

The developments introduced along this research enhance autonomous inspection capabilities in harsh sub-sea environments. MARESyE proves suitable for eye-in-hand inspections, enabling millimeter-precise structural assessments through on-site information fusion at the edge. This improved tri-dimensional scene understanding advances the state-of-the-art in autonomous offshore inspection operations, contributing to safer and more efficient infrastructure maintenance.

Resumo

A energia eólica em alto mar tem-se destacado como uma tecnologia renovável cada vez mais relevante devido a fatores de capacidade superiores e turbinas de maior dimensão comparativamente às instalações em terra, exigindo procedimentos de manutenção eficazes em ambientes marítimos adversos. Isto resulta na necessidade de inspeções cíclicas para avaliar a integridade estrutural e detetar possíveis defeitos em componentes submersos. Embora as soluções robóticas ofereçam alternativas mais seguras às operações manuais que expõem mergulhadores e pessoal de apoio a riscos significativos, as capacidades percetivas integradas nestes sistemas proporcionam representações uni-modais degradadas, limitadas pelo alcance de aquisição de dados e oclusões. Atualmente, os sensores de imagem para ambientes submersos carecem de estratégias de mitigação para reduzir o impacto da absorção de luz, dispersão e partículas em suspensão. Estes não só reduzem a visibilidade, limitam o contraste e introduzem neblina, como também dificultam criticamente a aquisição de informação tri-dimensional de alta fidelidade, crucial para a avaliação estrutural autónoma.

Este doutoramento centra-se no desenvolvimento de soluções de imagem subaquática adaptadas para perceção eye-in-hand. O sensor de imagem híbrido MARESy é desenvolvido para satisfazer requisitos de inspeção de proximidade, recolhendo informação com precisão milimétrica de alta fidelidade através de imagem stereo passiva e triangulação laser ativa. Um sistema de aquisição range-gated orientado aos dados obtidos, mitiga o impacto de fenómenos submarinos, enquanto princípios de fusão de informação são introduzidos para aproveitar as características complementares de cada modalidade. Esta abordagem permite a integração de dados texturados, nuvens de pontos tri-dimensionais e princípios de perceção de movimento em representações unificadas do cenário submarino.

O sensor MARESy demonstra a sua eficácia através de processamento embebido e medições milimétricas dentro do seu alcance operacional ótimo, mantendo um desempenho fiável no Centro de Testes ATLANTIS apesar de desafios como agregação de organismos, variações de iluminação e partículas em suspensão. A informação tri-dimensional baseada em visão recolhida em ambiente submarino exhibe perfis de erro e características de dados distintos: stereo fornece informação densa e texturada mas o seu desempenho métrico degrada-se com a distância, enquanto a triangulação laser ativa estima um conjunto esparso de medições precisas com cobertura espacial limitada. São introduzidas metodologias de fusão inovadoras para explorar estas características complementares, representando as primeiras abordagens à integração de dados 3D heterogéneos de técnicas de imagem subaquática. O algoritmo Joint Masked Regression (JMR) propaga medições baseadas em laser através de curvas de regressão em regiões homogéneas, enquanto a *underwater Heterogeneous 3d information fusion Architecture* (RHEA) extrai conjuntamente características de nuvens de pontos acopladas através de uma abordagem de treino synthetic-to-real. Ambas as metodologias alcançam melhorias significativas em ensaios controlados, com ganhos de precisão de 34,9% e 52,5% sobre reconstruções stereo tomadas como base de comparação, mantendo o desempenho em condições próximas do real no Centro de Testes ATLANTIS. A estimação de

profundidade baseada em stereo permanece fundamentalmente limitada por ambiguidades na correspondência de características e degradação contínua de precisão. Esta tese aborda estas limitações através de princípios de fusão multimodal, introduzindo a arquitetura fusiNg tEXture with apparent motion information for underwater Scene reCOstruction (NESO) para obter representações densas e texturadas através da fusão de dados de imagem com princípios de percepção de movimento aparente. A rede utiliza um design piramidal baseado em atenção que preserva características de alto nível ao reintroduzi-las ao longo de múltiplos passos de upsampling. A validação com inspeções de ânodos sacrificiais demonstra erros absolutos até 0,0016 m, embora as previsões requeiram refinamento fora do intervalo métrico de treino. A integração com dados laser esparsos aborda esta limitação, alcançando melhorias métricas até 0,1124 m enquanto preserva relações de profundidade relativa.

Os desenvolvimentos introduzidos ao longo deste doutoramento melhoram as capacidades de inspeção autónoma em ambientes submarinos adversos. O sensor MARESyE prova-se adequado para inspeções eye-in-hand, permitindo avaliações estruturais milimétricas através de fusão de informação no local. Esta compreensão tri-dimensional melhorada avança o estado da arte em operações de inspeção em alto mar autónomas, contribuindo para uma manutenção de infraestruturas mais segura e eficiente.

Acknowledgements

While this work represents years of my personal dedication and sacrifice, it would not have been possible without the amazing people that support, guide, and love me every day.

My supervisor, Andry Maykol Pinto, has been my scientific mentor since I first ventured into research, instilling in me his commitment to fair, transparent, and exceptional research. Throughout this journey, he supported me through personal challenges and scientific breakthroughs. I'm deeply proud to be part of his team of tenacious and driven individuals who continue to push boundaries and advance maritime robotics to unprecedented heights. Professor Andry's values and dedication are reflected not only in his own achievements but in the success of his entire team. I am grateful for the invaluable opportunities he provided, which enabled me to grow and meet his high standards of excellence. Above all, I thank Andry for his constant dedication, helping me become a better version of myself, both professionally and personally.

I believe Daniel Campos represents the invaluable piece every team needs. Daniel took me in from day one, endured my constant interruptions and attempts to destroy his computer. He became more than a colleague, a big brother who I have turned to for help in and outside of work. I thank him for all the time and effort he dedicated to me, for his endless supply of dad jokes (which existed even before his beautiful baby girl), and for his willingness to share his experience and knowledge. I would also like to thank Rafael Claro for his readiness to help, whether devising his MacGyver-worthy solutions, finding the most obscure suppliers, or creating his memorable (if questionable) acronyms. My journey would not have been complete without the invaluable contributions of my mechanical engineers Pedro Pereira and João Dionísio. They supported me throughout every phase, from design and conceptualization to experimental testing, often sacrificing their own personal time and research priorities to advance my work. Hugo Campos is one of the funniest people I've met, combining an incredible drive with an infectious sense of humor that made every workday brighter. I extend my gratitude to Lourenço Pinho, our lovable Lupinho, for always spending his days (or half-days) striving to make me laugh, feeding me sugar bombs, and cursing seagulls. I am also grateful to my MOTUS colleagues João Marques, Francisco Neves, Celso Pereira, Luís Branco and João Campanhã for all the comradeship moments and brainstorm sessions. A special thanks to José Carlos Azevedo for his expert guidance in electronics systems design.

To my dear friend Renato Silva, who has walked this path beside me every step of the way: thank you for always backing me up and pushing me through. We have shared a decade filled with laughter, challenges, triumphs, and setbacks. From our first conference adventure to our FCT morning celebratory beer, and now to this milestone we've fought so hard to achieve. I feel quite fortunate to have friends who have become family, and David Lamas is certainly one of those friends. Without his belief in me, this journey might never have begun, as he was instrumental to kickstart my PhD venture. More than a thank you is owed to my friend Daniel Barbosa, who has been my compass and voice of reason, the one I turn to for wisdom across countless subjects, and who reminds me, with unwavering certainty, to "trust the process". Tiago Sousa has the ability

to find just the right words when I need them most. His understanding of my struggles and his constant efforts to lift my spirits have been invaluable throughout this journey. My gratitude for Isabel Geriante runs as deep as the times we've slipped and fell together. She has been both a cheerleader and catalyst in my journey, giving me the confidence to take risks and pursue what makes life truly meaningful. I am also grateful to Diogo Duarte, Bruno Miranda, Joana Gomes, Renato Cruz, and Vitor Carneiro for our years of friendship, shared memories, and support.

During this adventure, I met an incredible person who began as the coolest Bambi and sassiest co-worker, grew into a dear friend, and became the person that I love, cherish, and admire. My gorgeous girlfriend Inês Pereira has helped grow so much in such a short time, that I find it hard to describe how fortunate I feel to have found such a deep and loving connection. I am thankful for your constant support, words of encouragement, and for making me believe it's possible. I cannot express how grateful I am to have someone like you to share my victories with. I love you.

To my grandpa Manuel Osório who is the man I model myself after. I know you would be smiling ear-to-ear seeing me finish this journey. My gratitude is immense for all you have done for me, and I know that even if your mind wandered, your heart always followed me everywhere. I will always stay true to the values you instilled in me. My life has always been guided by strong women, starting with my dear grandmas, Ermelinda and Beatriz. A most sincere thank you for shaping me into the man I am today. Finally, to my mother, who sacrificed everything for me to be where I am today. I can never fully express my gratitude for your boundless love and support. You have always believed in me and encouraged me to push further and follow my gut, even when it meant facing greater challenges and hardships. You are the backbone of my achievements, please know that I am deeply proud of the warrior mother you are.

Pedro Nuno Barbosa Leite

Official Acknowledgements

Pedro Nuno Barbosa Leite acknowledges the Portuguese Government through the FCT (Foundation for Science and Technology) for the PhD grant 2020.06949.BD (<https://doi.org/10.54499/2020.06949.BD>). This work is also funded by the European Union's Horizon Europe research and innovation programme, under the Grant Agreement No. 101189723 (AEROSUB). Acknowledgments are further extended to the Exail team: Fabrice Calderoni, Jonathan Prevosto, Nicolas Bernard and François Chermeux; for their continuous support and invaluable experience during the ATLANTIS Test Center experiments.



*“Strive for perfection in everything you do.
Take the best that exists and make it better.
When it does not exist, design it.”*

Sir Henry Royce

Contents

1	Introduction	1
1.1	Motivation	3
1.2	Objectives	6
1.3	Document Structure	6
1.4	Contributions	8
1.4.1	Achievements	8
1.4.2	Publications	9
2	Literature Review	11
2.1	Introduction	11
2.2	The Challenges of Underwater Perception	13
2.2.1	Sub-Sea Physical Phenomena	13
2.2.2	Calibration of Underwater Optical Systems	16
2.3	Underwater Tri-dimensional Scene Understanding	18
2.3.1	Reconstruction Techniques and Imaging Sensors	18
2.3.2	Perception-based Information Fusion	31
2.3.3	Eye-in-Hand Perception	38
2.4	Critical Analysis	41
3	MARESyE: Hybrid Imaging for Tri-dimensional Underwater Scene Understanding	45
3.1	Introduction	45
3.2	Hybrid Imaging and Data-driven Formulation	47
3.3	Optical Design for High-quality Underwater Imaging	49
3.4	Mechanical Design and Validation	52
3.5	Results and Discussion	55
3.5.1	Accuracy of the 3D Measurements	55
3.5.2	Inspecting Submerged Sacrificial Anodes	64
3.6	Final Considerations	73
4	Tri-dimensional Information Fusion for Inspecting Submerged Critical Structures	75
4.1	Introduction	75
4.2	Joint Masked Regression Algorithm	76
4.3	Underwater Heterogeneous 3D Information Fusion Architecture	78
4.4	Results and Discussion	82
4.4.1	Synthetic Data Experiments	83
4.4.2	Controlled Underwater Experiments	86
4.4.3	ATLANTIS Coastal Testbed - Real Application Experiments	88
4.5	Final Considerations	94

5	Fusing Texture, Motion Perception and Sparse Depth for 3D Scene Reconstruction	95
5.1	Introduction	95
5.2	Fusing Texture and Apparent Motion Components for Depth Estimation	96
5.3	Results and Discussion	101
5.3.1	NESO State-of-the-Art Comparison	102
5.3.2	Controlled Underwater Experiments	106
5.4	Final Considerations	110
6	Conclusion	111
6.1	Future Work	113
	References	115

List of Figures

1.1	Information fusion in underwater imaging enables enhanced perception capabilities. When integrated with eye-in-hand configurations, this approach facilitates autonomous inspection tasks crucial for offshore O&M procedures.	5
2.1	Simulation of a light wave traveling through water at different depths.	14
2.2	The impact of backscattering during image acquisition, as captured by an off-the-shelf imaging sensor.	15
2.3	A visual representation of the two main sub-sea phenomena affecting underwater optical systems.	16
2.4	The refraction of light caused by the interactions between two mediums, for distinct viewport geometries.	17
2.5	The cylinder-shaped AQUASENSOR houses an off-the-shelf stereo camera used by divers for reconstructing and mapping corals and shipwrecks.	19
2.6	The Fugu-f optical system is designed to be a flexible sensor capable of retrieving high-resolution tri-dimensional information.	20
2.7	A low-cost stereo vision system developed from off-the-shelf technology for fish length measurement.	21
2.8	Laser beam projections on the scene with known patterns that allow for the scale of the SfM reconstruction to be estimated - figure a). Final tri-dimensional reconstruction is depicted in figure b).	23
2.9	A single laser beam is diffracted into multiple lines, with lower-powered diffracted lines attenuating first as turbidity levels rise.	24
2.10	Gray-coded pattern projector improves the matching of a stereo composition. As turbidity levels increase, the quality of the estimated reconstructions quickly degrades.	25
2.11	A galvanometer iteratively redirects the projected laser beam allowing for increased spatial coverage of the scenario.	26
2.12	Concept design of the MARESyE hybrid imaging sensor: this system integrates both passive and active imaging techniques for underwater applications.	27
2.13	Hybrid imaging system composed of a stereo camera pair and a laser diode mounted on a sliding platform.	28
2.14	The UW-GAN is composed of two sub networks: UWC-Net estimates a coarse depth map; and UWF-Net refines this representation by combining it with the initial RGB input image.	29
2.15	The WaterMono training pipeline involves two stages: first, self-supervised training for teacher networks to generate pseudo depth labels, followed by training of student networks using a mix of self-supervised and supervised methods.	30

2.16	Sparse laser triangulation data is combined with surface normals from stereo methods, minimizing depth and normal vector errors with adaptively calculated weights.	33
2.17	A configuration of two orthogonal multibeam sonars provide horizontal and vertical perspectives. Features from the overlapping regions can be matched to obtain an improved 3D representation of the target object.	34
2.18	Self-supervised network that combines RGB images with masked echosounder cone measurements.	37
2.19	The SAUVIM AUV equipped with a 7-DOF MARIS 7080 manipulator uses an eye-in-hand monocular camera for object detection and tracking during a search and recovery mission.	39
2.20	A hybrid eye-in-hand configuration. The laser diode is attached to the manipulator, while a monocular camera gathers image data from the AUV's perspective. . . .	39
2.21	Custom end-effector gripper and camera to enable visual pipe detection, segmentation, and grasping point estimation from an eye-in-hand perspective.	41
2.22	Hybrid visual servoing configuration for underwater vehicle-manipulator systems, combining an ROV-mounted stereo camera and an eye-in-hand camera.	41
3.1	The MARESyE hybrid imaging sensor.	46
3.2	The photogrammetric stereo stage is depicted in figure a). Figure b) portrays the light stripe ranging acquisition. The heterogeneous tri-dimensional information gathered by the MARESyE sensor is illustrated in figure c).	47
3.3	A range-gated synchronizer is employed by MARESyE to mitigate the impact of the floating particles.	48
3.4	Internal design implications on the optical setup of the MARESyE sensor.	51
3.5	FEA stress results obtained from applying a uniform pressure of 3 MPa to the MARESyE housing structure.	53
3.6	FEA displacement results obtained from applying a uniform pressure of 3 MPa to the MARESyE housing structure.	54
3.7	Response of the MARESyE housing over time during the hyperbaric chamber experiment.	54
3.8	First experimental setup: assess the overall accuracy of MARESyE's 3D measurements by collecting PS and LSR point clouds at increasingly larger distances from a known target surface.	57
3.9	Evolution of the mean absolute error in the PS and LSR global depth estimations according to the distance from the MARESyE sensor.	58
3.10	A study on the flatness of MARESyE's PS point clouds was conducted for three distinct metric ranges.	59
3.11	Second experimental setup: analyze MARESyE's PS and LSR relative depth estimation accuracy for increasingly larger distances from the sensor.	60
3.12	Evolution of the mean absolute error in the PS and LSR relative depth estimations according to the distance from the MARESyE sensor.	61
3.13	Impact of the distance from the MARESyE sensor in the LSR relative depth estimation.	62
3.14	Evolution of the mean absolute error in the PS and LSR spatial dimension estimation according to the distance from the MARESyE sensor.	64
3.15	Decommissioned sacrificial anode cut off from the DURIUS floating structure used during the controlled underwater experiments.	65
3.16	MARESyE 3D reconstructions within a controlled indoor pool scenario.	66
3.17	Set of underwater experiments conducted at the ATLANTIS Coastal Test Center.	68

3.18	Sacrificial anode from the DURIUS structure used as the reconstruction target. . .	69
3.19	MARESyE 3D reconstructions within a real scenario at the ATLANTIS Test Center. . .	71
4.1	Visual representation of the joint masked regression algorithm.	77
4.2	Visual representation of the RHEA network.	79
4.3	Visual comparison between the proposed fusion methodologies using synthetic information.	84
4.4	RMSE maps obtained by comparing the respective Predictions a), b) and c). . . .	85
4.5	Set of objects used during controlled underwater experiments.	86
4.6	Comparison between the distinct point cloud predictions obtained during trials with object A.	88
4.7	INESC TEC team gathering underwater information using MARESyE suspended from a support vessel at the ATLANTIS Coastal Testbed in Viana do Castelo. . .	89
4.8	Comparison between the distinct point cloud predictions obtained during trials with object D, at the ATLANTIS Coastal Testbed.	90
4.9	Visual comparison between the proposed fusion methodologies using real underwater data collected by the MARESyE sensor at the ATLANTIS Coastal Testbed. . .	93
5.1	Depth information can be inferred from pixel displacements between two simultaneously captured viewpoints.	99
5.2	The NESO network follows an encoder–decoder architecture to combine RGB data with motion perception information to obtain a dense tri-dimensional representation of the underwater scenario.	100
5.3	Qualitative comparison of depth predictions on FLSea-Stereo dataset samples. . .	104
5.4	Root mean squared error maps comparing network predictions with ground truth depth measurements.	105
5.5	NESO and UW_Depth depth map predictions for the top and bottom sections of a sacrificial anode, inferred during a close-range inspection procedure conducted within controlled underwater environments.	107
5.6	Sacrificial anode reconstructions obtained by fusing sparse LSR information with the dense NESO point cloud prediction, using the JMR algorithm.	109

List of Tables

2.1	Advantages and disadvantages of distinct sensors and techniques used for underwater tri-dimensional perception.	43
3.1	Evolution of MARESy's PS point cloud spatial resolution in function of Z. . . .	50
3.2	Evolution of the uncertainty in MARESy's PS depth estimation in function of Z. . .	52
3.3	SGBM parameters used to estimate the PS point clouds for every experiment. . .	56
3.4	MARESy global depth characterization - PS and LSR measurements.	58
3.5	MARESy relative depth and spatial dimensions characterization - photogrammetric stereo measurements.	63
3.6	MARESy relative depth and spatial dimensions characterization - light stripe ranging 3D measurements.	63
3.7	Analysis of MARESy's relative depth measurements in the 3D reconstruction of a sacrificial anode within a controlled underwater environment.	66
3.8	Analysis of MARESy's photogrammetric stereo spatial dimensions measurements in the 3D reconstruction of a sacrificial anode within a controlled underwater environment.	67
3.9	Analysis of MARESy's relative depth measurements in the 3D reconstruction of DURIUS' sacrificial anode within a real underwater environment at the ATLANTIS Test Center.	70
4.1	Performance of the proposed fusion methodologies with synthetic data, compared to the estimated baseline.	84
4.2	Performance comparison between the proposed fusion methodologies using the underwater set of objects.	87
4.3	Performance comparison between the proposed fusion methodologies in a real underwater environment, using object D.	89
5.1	State-of-the-Art comparison on the FLSea-Stereo dataset.	103
5.2	Volumetric assessment of the anode reconstructions estimated by both the NESO and UW_Depth networks.	107
5.3	Volumetric assessment of the anode reconstruction obtained by fusing NESO's prediction with sparse LSR data through the JMR algorithm.	109

Abbreviations

2D	Two-Dimensional
3D	Tri-Dimensional
4D	Four-Dimensional
ACFM	Alternating Current Field Measurements
AI	Artificial Intelligence
AUV	Autonomous Underwater Vehicles
AVT	Allied Vision Technologies
CAD	Computer-Aided Design
CBAM	Convolutional Block Attention Module
CNC	Computer Numerical Control
CNN	Convolutional Neural Network
CTV	Crew Transfer Vessel
DOF	Degrees-Of-Freedom
DVL	Doppler Velocity Log
EC	Eddy Current
EM	Expectation-Maximization
FEA	Finite Element Analysis
FOV	Field-Of-View
FSS	Forward Scan Sonar
GPU	Graphics Processing Unit
HSE	Health Safety and Environmental
HSV	Hue, Saturation, Value
ICP	Iterative Closest Point
IMU	Inertial Measurement Unit
IPCC	Intergovernmental Panel on Climate Change
IR	Infrared
IRIS	Invariant 3D Reconstruction by Instrumented Stereovision
JMR	Joint Masked Regression
LCOE	Levelized Cost Of Energy
LED	Light-Emitting Diode
LiDAR	Light Detection and Ranging
LSR	Light Stripe Ranging
LTS-RA	Least Trimmed Square Rotation Averaging
MAE	Mean Absolute Error
MPI	Magnetic Particle Inspection
MRE	Mean Relative Error
NESO	fusiNg tExture with apparent motion information for underwater Scene recOn- struction

O&M	Operation and Maintenance
PS	Photogrammetric Stereo
PSMNet	Pyramid Stereo Matching Network
PSPNet	Pyramid Scene Parsing Network
PSNR	Peak Signal-to-Noise Ratio
RADAR	Radio Detection And Ranging
RAFT	Recurrent All-Pairs Field Transforms
RANSAC	RANdom SAMple Consensus
RGB	Red, Green, Blue
RGB-D	Red, Green, Blue, Depth
RHEA	undeRwater HEterogenous 3d information fusion Architecture
RMSE	Root Mean Squared Error
ROV	Remotely Operated Vehicle
RSfM	Refractive Structure-from-Motion
SAN	Style Adaptation Network
SGBM	Semi-Global Block Matching
SGD	Stochastic Gradient Descent
SIFT	Scale Invariant Feature Transform
SfM	Structure-from-Motion
SLAM	Simultaneous Localization and Mapping
SNR	Signal-to-Noise Ratio
SOCA-CFAR	Smallest Of Cell-Averaging Constant False Alarm Rate
SQUID	Stereo Quantitative Underwater Image Dataset
SSIM	Structural Similarity Index Measure
STD	STandard Deviation
UW-GAN	Underwater Generative Adversarial Network
YOLO	You-Only-Look-Once

Chapter 1

Introduction

The depletion of fossil fuels, coupled with increasing public environmental awareness and the pressing need to address climate change, has fostered a rapid transition towards green energy sources. The Intergovernmental Panel on Climate Change (IPCC) has established a critical limit of 1.5 °C for global temperature rise [1]. Meeting this target is only possible if renewable sources supply 75-85% of global electricity production [2]. Wind energy has emerged as a leading renewable option due to its technological maturity and cost-effectiveness [3]. The sector reached a global wind power capacity of 1 TW in 2023; however, current installation rates of 117 GW annually must accelerate to 320 GW by 2030 to meet global climate targets¹.

Offshore wind energy benefits from higher and more consistent wind speeds at sea, resulting in expected capacity factors of 50%, compared to 30-45% for onshore installations [4]. Furthermore, current offshore technology enables turbines that generate up to 15 MW of power, representing an increase of over 70% compared to typical land-based units [5]. The global offshore market has experienced notable growth, with installed capacity surpassing 64 GW in 2023, and future projections indicate an addition of 138 GW worldwide between 2024 and 2028 - a compound annual growth rate of 28%¹. This trend, driven by technological advancements and pioneering research projects like ATLANTIS [6] and EU-SCORES², is expected to accelerate, with total global capacity projected to reach 370 GW by 2030 and 630 GW by 2050³. While the technology maturity of bottom-fixed shallow water offshore platforms is superior to that of floating deep sea platforms, the latter offers compelling advantages including enhanced wind resources, site location flexibility, reduced transmission and distribution losses, and minimal visual and noise impact [7, 8]. Although floating offshore wind farms face higher deployment and operation costs [9], their viability is demonstrated by successful projects such as WindFloat Atlantic⁴, which has

¹Global Wind Energy Council, 2024. *Global Wind Report 2024*. URL: https://gwec.net/wp-content/uploads/2024/04/GWR-2024_digital-version_final-1.pdf. Accessed in November 2024.

²EU-SCORES. URL: <https://euscores.eu/>. Accessed in December 2024.

³C. Emanuel, 2023. *Offshore wind industry faces challenges and opportunities as global growth continues*. URL: <https://www.4coffshore.com/news/offshore-wind-industry-faceschallenges-and-opportunities-as-global-growth-continues-nid28083.html>. Accessed in November 2024.

⁴Windplus. *World's first semi-submersible floating offshore wind farm*. URL: <https://www.windfloat-atlantic.com/>. Accessed in November 2024.

been operating at 25 MW capacity since April 2020 off the Portuguese coast.

These infrastructures require regular inspection and maintenance to ensure sustained power production and maximize operational lifespan. The harsh offshore environment subjects wind turbines to continuous strain and fatigue [10]. Environmental factors such as wind, rain, and ice accumulation contribute to the need of cyclical inspections of critical components, both above water (eg. blades, towers, transition pieces) and below the surface (eg. catenary and mooring lines, cathodic protection systems, scour protection layers) [11, 12]. Operation and Maintenance (O&M) for offshore wind farms costs 4-5 times more than onshore installations, accounting for 23% of total life-cycle costs [7]. The offshore wind O&M market is projected to experience substantial growth, with a 16% annual increase reaching 10 billion euros by 2029⁵. This market expansion is expected to drive innovation and efficiency improvements, potentially reducing the Levelized Cost Of Energy (LCOE) by over 10% by 2025 and one-third by 2035⁶. Current O&M methodologies rely heavily on manual operations, requiring divers, large vessels, support personnel, and specialized equipment. Despite strict Health Safety and Environmental (HSE) regulations, human operators conducting these inspection activities remain exposed to significant risks [13]. Weather limitations, including maximum wave heights of 1.5 m and wind speeds of 10 ms^{-1} , restrict the yearly operational window to less than 40% [14]. Major operational costs arise from the daily use of Crew Transfer Vessels (CTVs), personnel expenses and revenue losses due to turbine downtime during inspection procedures [15]. The integration of autonomous solutions into O&M operations can yield cost savings exceeding 20% of total wind farm investment [16]. These systems significantly enhance safety by eliminating human exposure to hazardous operations, particularly for underwater inspections where divers traditionally perform visual and non-destructive testing. Autonomous platforms provide systematic, repeatable inspection patterns and high-resolution data that enable detailed structural assessment, overcoming reliability issues and human errors inherent to diver-based operations [17]. Their ability to function in harsh conditions while providing real-time data expands the O&M operation window, allowing proactive maintenance strategies for early defect and fault detection [18]. This shift towards autonomous inspection ultimately leads to a more cost-effective O&M strategy, reducing wind energy production downtimes and consequently decreasing the overall LCOE.

The adoption of Autonomous Underwater Vehicles (AUVs) and Remotely Operated Vehicles (ROVs) for sub-sea inspection and intervention tasks has grown significantly, fostered by large scale research initiatives (eg. ATLANTIS [6], ORCA HUB [19]). These robotic platforms can be equipped with state-of-the-art sensors to retrieve real-time Two-Dimensional (2D) and Tri-Dimensional (3D) information for visual and volumetric assessments, while specialized tools and crawlers enable effective light intervention operations [20]. ROVs are the primary solution for

⁵Wood Mackenzie, 2021. *Global offshore wind O&M market to grow 16% annually to US\$12 billion by 2029*. URL: [https://www.woodmac.com/press-releases/global-offshore-wind-om-market-to-grow-16-annually-to-us\\$12-billion-by-2029/](https://www.woodmac.com/press-releases/global-offshore-wind-om-market-to-grow-16-annually-to-us$12-billion-by-2029/). Accessed in November 2024.

⁶A. Vasdev, Energy Economics, BloombergNEF, 2023. *2H 2023 LCOE Update: An Uneven Recovery*. URL: <https://about.bnef.com/blog/2h-2023-lcoe-update-an-uneven-recovery/>. Accessed in November 2024.

deep-sea search and rescue missions⁷, while also enabling survey operations, pipeline inspections, and marine growth removal [21, 22]. However, the attached tether significantly constrains vehicle maneuverability and operational range. ROV operations incur substantial costs, requiring support vessels, dedicated operators, and crew members for deployment and tether management [23]. In contrast, AUVs are designed for autonomous operation, with demonstrated capabilities in broad area surveys and acoustic inspections of submerged structures [24, 25]. Despite their advantages, AUV technology lacks the maturity required for intricate offshore tasks such as close-range inspections, biofouling removal, and intervention procedures [23]. Nevertheless, successful trials in controlled environments give insight into the technology's future potential [26]. As these platforms incorporate advanced Artificial Intelligence (AI)-driven perception and decision-making capabilities, autonomous vessels will significantly impact the O&M market, eliminating human operator dependencies and consequent fatigue-induced performance degradation.

1.1 Motivation

The most significant limitation of underwater vehicles remains their inability to obtain a comprehensive tri-dimensional perception of the surrounding environment [27]. Remote sensing devices commonly employed in underwater ventures retrieve either low-resolution acoustic data or degraded visual information, severely impacting critical operations including close-range navigation, visual fault detection, millimeter-precise volumetric assessments, and reliable robotic interventions [28]. Sonar technology, despite its robustness to water conditions and long-range capabilities, faces fundamental limitations for inspection tasks: high minimum operation distances, poor performance with complex geometries, and susceptibility to multipath reflections [27]. These limitations are further compounded by spatial resolution degradation with distance due to beam divergence [29]. Optical sensors offer an alternative approach, providing dense, high-resolution data with color and texture information crucial for detailed structural assessment [30]. However, sub-sea phenomena severely impact the acquisition of high-fidelity visual information. Light absorption underwater limits the effective visibility range, inducing the need for external illumination sources that often lead to inconsistent scene exposure. Moreover, suspended particles increase water turbidity and cause light scattering, where photons are reflected back to the imaging sensor, resulting in degraded images with reduced contrast, blur, and color distortion [31]. Addressing underwater phenomena and its effects on data degradation is a pressing matter for enabling widespread deployment of autonomous solutions in the O&M market. Visual inspections in underwater scenarios typically rely on off-the-shelf imaging technology integrated as payload within underwater vehicles. The dependence on a single source of information, coupled with the lack of mitigation strategies, makes these conventional sensors particularly vulnerable to sub-sea phenomena [30].

⁷Pelagic Research Services, 2024. URL: <https://www.youtube.com/watch?v=NhLcnKktu9k>. Accessed in November 2024.

Intervention missions can be conducted by sub-sea vessels equipped with robotic manipulators. However, successful deployment of these robot-manipulator systems currently requires human supervision, extensive area surveys, and *a priori* knowledge to identify target objects [32]. Traditional setups employ complex algorithms to handle occlusions caused by the robotic arm itself, as operations rely on vessel-mounted imaging sensors [33]. Eye-in-hand configurations address these limitations by integrating sensing capabilities directly on the robotic manipulators. This approach enables dynamic viewpoint adjustment without compromising vessel safety, while maintaining fixed relative distances to target surfaces. The resulting close-range perspective allows for high-fidelity information retrieval from otherwise inaccessible viewpoints, enabling precise structural assessment [34]. The enhanced volumetric information is crucial for informed decision-making during intervention operations, which currently lack the millimeter-precise tri-dimensional data necessary for accurate target pose estimation and grasp planning [35]. Despite their potential to significantly enhance the perceptual capabilities of autonomous O&M robots, current eye-in-hand imaging configurations are largely limited to monocular visual feedback or basic structured light reconstructions [36, 37].

The research presented in this doctoral thesis aims to improve the tri-dimensional scene understanding of submerged O&M vehicles by developing underwater imaging principles that enable the capture of high Signal-to-Noise Ratio (SNR) information from sub-sea environments. By diverging from traditional approaches, this work addresses the challenges posed by marine snow to gather high-fidelity multimodal data, including textured images, multiple sources of tri-dimensional data, and the magnitude of apparent pixel motion. The improved perceptual capabilities enable operations currently limited by insufficient situational awareness, particularly the inspection of submerged critical structures requiring millimeter-precise volumetric assessment. The development of a self-reliant underwater imaging sensor, capable of internal data processing, significantly reduces inspection times by providing real-time structural assessment feedback. Furthermore, a compact and versatile design enables new eye-in-hand configurations that leverage the robotic arm's degrees of freedom to gather valuable information from multiple perspectives, enhancing the reliability of autonomous interventions in complex underwater environments.

Information gathered underwater provides data with complementary characteristics that can be fused into a comprehensive tri-dimensional representation of the submerged scenario. Image data captures crucial texture-based features and spatial scene relationships, though it alone cannot provide absolute depth measurements. Tri-dimensional point clouds capture depth information through different approaches: passive stereo imaging generates dense, textured reconstructions offering broad scene coverage but with accuracy degrading beyond calibration range due to its limited resolution; while active laser-based triangulation estimate a sparse set of millimeter-accurate measurements that provide valuable volumetric information with limited spatial coverage [38]. Additionally, motion-based depth estimation leverages apparent pixel displacements to infer pixel-to-pixel relative depth relationships, however these cannot be directly translated to metric measurements. Fusing information from multiple modalities enables more complete and robust perception of the environment by combining individual strengths while compensating for specific

limitations [39]. However, integrating heterogeneous data streams presents fundamental challenges: different sampling rates and spatial resolutions for each modality; measurements exhibit distinct error distributions and uncertainty profiles; and data formats vary in density, precision, and coverage patterns. Moreover, successful fusion requires precise temporal synchronization between data streams and accurate spatial alignment of their coordinate frames. This integration can be achieved through conventional probabilistic models, vision-based approaches, or deep learning architectures that automatically extract features from multimodal data [40]. The research documented in this thesis addresses the fusion of distinct data modalities to obtain a unified, dense, and precise tri-dimensional representation of the submerged scene. Novel information fusion algorithms are introduced to combine heterogeneous 3D data streams, mitigating the inherent limitations of single-modality depth estimation by propagating the accuracy of sparse structured light measurements throughout dense stereo-based reconstructions. Furthermore, this work develops multimodal fusion principles that combine high-fidelity textured, tri-dimensional, and apparent pixel displacement information in a multi-stage approach, leveraging motion perception to achieve dense and accurate depth estimation.

The integration of fusion algorithms within an edge-oriented underwater imaging sensor enables real-time processing of multimodal data, providing dense and accurate tri-dimensional representations. When coupled with eye-in-hand configurations, as illustrated in figure 1.1, this approach enables millimeter-precise inspection from multiple viewpoints, crucial for autonomous manipulation tasks. These enhanced perceptual capabilities have the potential to significantly reduce operational costs and increase the reliability of offshore O&M procedures through more efficient and automated inspection routines.

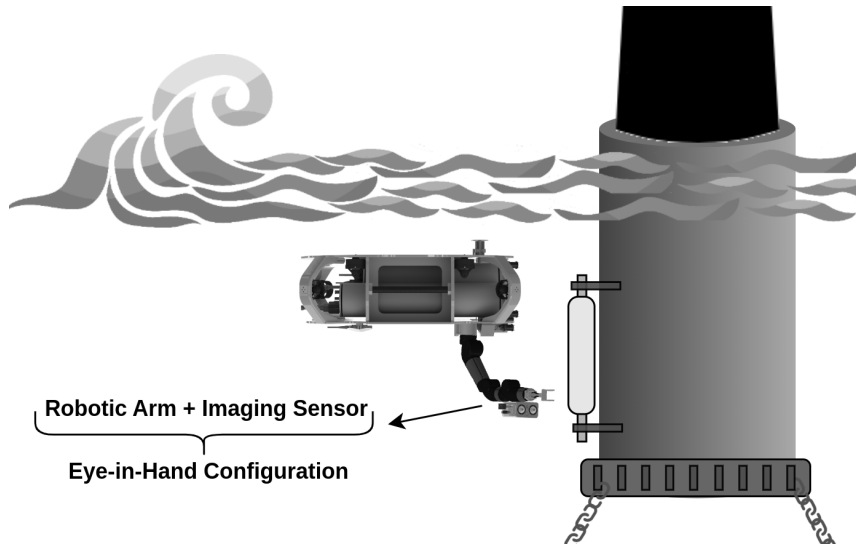


Figure 1.1: Information fusion in underwater imaging enables enhanced perception capabilities. When integrated with eye-in-hand configurations, this approach facilitates autonomous inspection tasks crucial for offshore O&M procedures.

1.2 Objectives

This research work aims to extend the perception capabilities of underwater vehicles through enhanced tri-dimensional scene understanding. The proposed approach develops an imaging sensor capable of retrieving high SNR data from both passive and active techniques (hybrid imaging), employing information fusion principles to combine diverse sensing modalities to obtain an improved tri-dimensional representation of the observed scenario. The objectives of this thesis are as follows:

- Obj1. Explore the advantages of fusing multimodal data streams obtained through data-driven hybrid imaging principles. Each modality contributes distinct information: image data encodes contextual cues and spatial relationships; tri-dimensional data ranges from high density textured point clouds to narrow sets of non-textured points with highly precise depth measurements; and motion analysis yields pixel-to-pixel relative depth cues. Vision and learning-based fusion approaches can integrate both modality-specific and shared features to leverage these complementary characteristics, creating unified representations that combine density and millimeter-precision for an enhanced volumetric inspection of submerged structures.
- Obj2. Design a hybrid underwater imaging sensor capable of retrieving high-fidelity multimodal information in harsh sub-sea environments by employing mitigation strategies that reduce data degradation caused by sub-sea phenomena (e.g. light absorption, scattering, turbidity). The sensor should maintain flexibility for different deployment configurations including vessel-mounted and eye-in-hand setups; while incorporating edge processing capabilities for real-time operation.
- Obj3. Validate the developed underwater imaging sensor and fusion methodologies through extensive experimentation in harsh near-real conditions at the ATLANTIS Coastal Testbed [41]. These trials will quantitatively and qualitatively assess the system's ability to maintain millimeter-precise measurements while enhancing point cloud density, evaluate its robustness to sub-sea challenges, and demonstrate its applicability for real offshore inspection tasks.

Therefore, this work introduces an underwater imaging solution that addresses critical problems that limit the widespread use of autonomous vehicles in the inspection and intervention O&M market. Providing enhanced tri-dimensional situational awareness to underwater vessels empowers safer and more reliable maintenance operations, reduces the need for on-site human intervention, and contributes to lowering operational costs in offshore infrastructure maintenance.

1.3 Document Structure

This document is organized as follows:

- Chapter 2 reviews the state-of-the-art in underwater scene understanding, starting with fundamental challenges of light-water interactions. The analysis covers various sensing approaches including photogrammetry, structured light principles, and learning-based solutions. Data fusion methodologies for combining heterogeneous tri-dimensional data and multimodal information are explored, along with specialized eye-in-hand configurations for close-range inspection. The chapter concludes with a critical discussion that identifies current limitations and research opportunities in the field.
- Chapter 3 details the development and validation of the MARESyE hybrid imaging sensor. The chapter presents both optical and mechanical design considerations, incorporating strategies to mitigate sub-sea phenomena while enabling high-resolution data acquisition through complementary passive and active imaging techniques. A thorough experimental validation is conducted, first in controlled environments to characterize the sensor's operational range and accuracy, then in real harsh conditions through integration with a commercial ROV at the ATLANTIS test site. The chapter demonstrates the sensor's effectiveness in reconstructing sacrificial anode structures despite challenging conditions.
- Chapter 4 presents novel methodologies for fusing heterogeneous tri-dimensional information for underwater structural inspection. Two distinct approaches are introduced: a regression-based algorithm that propagates sparse, accurate measurements to enhance dense point clouds, and a learning-based solution trained through a synthetic-to-real scheme that allows for direct underwater application. Both methods are extensively validated through simulated scenarios, controlled underwater trials, and real-world deployment, demonstrating their effectiveness in improving point cloud accuracy while maintaining density.
- Chapter 5 introduces a novel approach to underwater scene understanding through multi-modal fusion. The chapter introduces a learning-based architecture that combines textured images with apparent motion components to generate dense point clouds, which are further refined using precise sparse depth measurements. Comprehensive validation includes performance comparisons with state-of-the-art methods, generalization studies, and extensive testing from controlled environments to challenging real-world trials. The results demonstrate the system's enhanced perceptual capabilities for underwater inspection tasks through effective fusion of complementary data modalities.
- Chapter 6 synthesizes the major contributions and findings of this thesis, critically analyzing the novel aspects in underwater perception and their impact on offshore inspection capabilities. The chapter concludes by providing comprehensive guidelines for future research directions.

1.4 Contributions

This section outlines the key contributions and impact of this research work in advancing underwater perception capabilities. The sections below encompass the significant technical achievements, publications, and collaborative developments that have emerged from this thesis.

1.4.1 Achievements

The most relevant achievements of this thesis are:

1. Heterogeneous tri-dimensional fusion techniques that combine textured dense 3D information with a very accurate sparse set of laser-based points into a single representation. A first methodology expands the laser information towards neighboring homogeneous regions through a regression curve that characterizes displacements between both point clouds. The second principle employs a deep learning-based approach to fuse both point clouds through an early fusion encoder-decoder architecture that conjointly learns features from a coupled representation of both inputs. Extensive validation with real data demonstrates an average absolute error of 0.0133 m and 0.0097 m, corresponding to a 34.9% and 52.5% improvement, respectively, over the input measurements.
2. An encoder-decoder architecture that leverages multi-stage pyramidal attention-based skip connections to fuse image data with the apparent movement of pixels to estimate a dense 3D representation of the submerged scenario. The network bypasses the need for parametrization by learning directly from the coupled input data. Introducing apparent motion components into the model's optimization process incites an easier transition to new environments: the proposed architecture outperforms monocular-based depth estimation methods, estimating point clouds with relative errors down to 4.57%.
3. A synthetic-to-real supervised training procedure. The heterogeneous tri-dimensional data retrieved by MARESyE is faithfully emulated in simulation. This allows for vast quantities of synthetic training data; bypassing the lack of available 3D underwater datasets in the scientific community, and the consequent need for transfer learning techniques.
4. A new design of the MARESyE sensor capable of significantly enhancing the remote sensing capabilities of AUVs and ROVs by employing a data-driven formulation that mitigates photogrammetric sub-sea hindrances. The high-fidelity tri-dimensional information captured through the internal processing of both active and passive imaging techniques, makes the MARESyE sensor suitable for the underwater reconstruction and volumetric assessment of critical structural elements in offshore platforms.
5. A comprehensive understanding of the accuracy in the tri-dimensional information obtained from the MARESyE sensor. Consecutive experiments conducted in a controlled underwater setting allow for a thorough analysis of the fidelity of 3D measurements and their susceptibility to factors such as distance from the sensor, camera parameterization, and error

characteristics inherent to each technique. The gained insights reveal that the $[0.5, 0.6]$ m distance spectrum is the optimal operational range for the sensor, wherein both passive and active techniques consistently achieve millimetric accuracy.

6. Extensive qualitative and quantitative trials employ the MARESy sensor to reconstruct and assess the state of sacrificial anode structures. In controlled environments characterized by low turbidity and the absence of external hindrances, MARESy estimates a reconstruction of the target element with absolute errors down to 0.0008 m; a 1.10% error relative to the ground truth reading. The optical prowess and robust design (depth-rated for 200 m) of the MARESy sensor are further validated in a real underwater application where it enhances the remote sensing capabilities of a commercial ROV. The embedded processing of the high SNR information, captured in these harsh environments, facilitates on-site assessments of the structure in terms of anode wear, possible structural damages and the presence of biofouling.

1.4.2 Publications

Main publications that result from this research:

- [38] P. N. Leite and A. M. Pinto, "Fusing heterogeneous tri-dimensional information for reconstructing submerged structures in harsh sub-sea environments", *Information Fusion*, 103:102126, 2024. doi:10.1016/j.inffus.2023.102126.
- [31] P. N. Leite, P. N. Pereira, J. M. M. Dionísio, and A. M. Pinto, "Hybrid underwater imaging for the tri-dimensional inspection of critical structural elements in offshore platforms," *Ocean Engineering*, 314:119658, 2024. doi.org/10.1016/j.oceaneng.2024.119658.
- [42] P. N. Leite and A. M. Pinto, "Exploiting motion perception in depth estimation through a lightweight convolutional neural network", *IEEE Access*, 9:76056–76068, 2021. doi:10.1109/ACCESS.2021.3082697.
- [43] P. N. Leite, R. J. Silva, D. F. Campos, and A. M. Pinto, "Dense disparity maps from rgb and sparse depth information using deep regression models", in Aurélio Campilho, Fakhri Karray, and Zhou Wang (eds), *Image Analysis and Recognition*, pages 379–392, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-50347-5_33.

Supplementary publications that stem from this work:

- [44] J. M. M. Dionísio, P. N. A. A. S. Pereira, P. N. Leite, F. S. Neves, J. M. R. S. Tavares, and A. M. Pinto, "Nereon - an underwater dataset for monocular depth estimation," in *OCEANS 2023 - Limerick*, 2023, pp. 1–7. doi:10.1109/OCEANSLimerick52467.2023.10244675.
- [45] J. Mina, P. N. Leite, J. Carvalho, L. Pinho, E. P. Gonçalves, and A. M. Pinto, "Enhancing Underwater Inspection Capabilities: A Learning-Based Approach for Automated

Pipeline Visibility Assessment", in: L. Marques, C. Santos, J. L. Lima, D. Tardioli, M. Ferre (eds), Robot 2023: Sixth Iberian Robotics Conference. ROBOT 2023. Lecture Notes in Networks and Systems, vol 978. Springer, Cham. doi.org/10.1007/978-3-031-59167-9_14

- [46] J. Carvalho, P. N. Leite, J. Mina, L. Pinho, E. P. Gonçalves, and A. M. Pinto, "Artificial Intelligence for Automated Marine Growth Segmentation", in: L. Marques, C. Santos, J. L. Lima, D. Tardioli, M. Ferre (eds), Robot 2023: Sixth Iberian Robotics Conference. ROBOT 2023. Lecture Notes in Networks and Systems, vol 978. Springer, Cham. doi.org/10.1007/978-3-031-59167-9_13

Master's theses developed from this research:

- M. N. Rosendo, "Enhancing the Tri-Dimensional Awareness of an Autonomous Surface Vehicle through Machine Learning Algorithms", M.Sc. dissertation, Faculty of Engineering of the University of Porto, FEUP, 2022. hdl.handle.net/10216/140752.
- J. P. S. Esteves, "Tri-dimensional Reconstruction of Submerged Objects through an Active Eye-in-Hand Perception System", M.Sc. dissertation, Faculty of Engineering of the University of Porto, FEUP, 2022. hdl.handle.net/10216/144571.
- D. J. M. Silva, "Underwater Imaging and 3D Sensor Fusion", M.Sc. dissertation, Faculty of Engineering of the University of Porto, FEUP, 2024. hdl.handle.net/10216/160509.

Chapter 2

Literature Review

One of the major constraints of underwater perception is the lack of sensors capable of providing high-quality 3D information [47]. The acquisition of this information is hindered by numerous factors, e.g. the range and resolution of the sensor, poor illumination conditions or lack of texture within the scenario [43]. These issues become even more accentuated within sub-sea environments, as these aggravate data degradation. Section 2.2 discusses how the interactions of light underwater affect the quality of the information gathered by imaging sensors. A myriad of solutions have been explored to equip underwater vehicles with tri-dimensional perception capabilities. Section 2.3 explores various perception systems and imaging techniques, as well as methodologies for combining information from these complementary sources. Eye-in-hand configurations, where sensors are directly coupled to robotic manipulators, enable dynamic and detailed close-range perception of the scene. Section 2.3.3 reviews the literature on these specialized configurations and their applications in underwater environments. Finally, a critical analysis is presented in section 2.4, summarizing key insights and identifying limitations and research gaps in the current state of the art.

2.1 Introduction

Autonomous underwater vehicles and ROVs are currently unable to function fully autonomously in close-range navigation, fault inspection, target detection and intervention tasks; due to being limited when it comes to having a complete tri-dimensional perception of their surrounding space [38]. While surface and aerial state-of-the-art technologies make use of multimodal perception systems to navigate in cluttered scenarios [48], land aerial vehicles [49, 50], and inspect critical structures [51]; underwater vessels commonly resort to acoustic or vision uni-modality perception systems.

Sonar-based technology has been used in maritime applications for tasks such as bathymetry, navigation and obstacle avoidance [52, 53, 54]. Information is retrieved by emitting a pulse of sound and listening to its reflections. As sound propagates faster in water, the ability to take long range measurements is one of the main advantages of sonars. Acoustic-based data has also

been employed to obtain submerged tri-dimensional reconstructions. The work of T. Guerneve *et al.* (2018) [55] employs wide-aperture imaging sonars for underwater object reconstruction, using both a spatially varying blind deconvolution method and a nonlinear space carving approach. Field tests on a polygonal target reveal median errors of 0.112 m and 0.089 m, respectively. P. V. Teixeira *et al.* (2019) [56] propose a unified framework that simultaneously estimates multibeam sonar poses, measurements and model parameters for reconstructing underwater scenes. Both the sonar segmentation and point cloud reconstruction problems are formulated as factor graphs, and solved concurrently alongside a Simultaneous Localization and Mapping (SLAM) problem. Reconstructions estimated from real data show a mean spatial uncertainty of 0.059 m. A multi-view scanning method that leverages the tri-dimensional information retrieved from a Forward Scan Sonar (FSS) to choose the next scan path is proposed B. Kim *et al.* (2020) [57]. The Iterative Closest Point (ICP) algorithm is used for the quantitative analysis of field experiments conducted in a coastal area, revealing Root Mean Squared Errors (RMSE) as low as 0.046 m. The research introduced by M. Qadri *et al.* (2023) [58] represents data gathered from an imaging sonar as neural implicit functions. A differential volumetric renderer is then employed to derive sonar measurements. Experimental reconstructions from water tank trials show that a 28° sonar aperture can achieve RMSE as low as 0.072 m. Y. Feng *et al.* (2024) [59] present a differentiable space carving method that leverages occupancy probability grids and multi-resolution hash encoding to render echo probabilities instead of intensities. Experiments in a water tank display accurate measuring of spatial dimensions, but poor overall shape of the reconstructed object. The work of W. Xu *et al.* (2024) [60] details an inverse sensor model for forward-looking sonars that generates occupancy-elevation grid maps by extracting probability and elevation information from the captured image. Simulation results show height reconstruction errors ranging from 0.035 to 0.094 m. Close-range fine reconstructions with sonar information prove difficult due to overlapping echoes, lack of millimeter-precise accuracy and, the need for swiping motions to increase spatial coverage [27]. Additionally, the reliance on continuous movement can introduce further reconstruction errors, particularly due to navigation drifts [55].

Imaging sensors are versatile and capable of gathering rich visual-based information. These commonly resort to off-the-shelf stereoscopic systems in watertight enclosures, or single high resolution cameras coupled with laser or pattern projectors [61, 47, 62]. The reliance on a single source of information makes these sensors vulnerable to sub-sea phenomena [30]. Light absorption, scattering and the presence of suspended particles lead to extensive data degradation that difficulties the process of acquiring accurate and high resolution information. The implications of this noisy information reflect themselves directly on the autonomous capabilities of AUVs, since these become restricted of knowledge about their surroundings, limiting their maneuverability and operation [63]. Research to mitigate these photogrammetric limitations typically resorts to post-processing image enhancements or range-gated systems that filter some scattering components during image formation [64]. The distinct data modalities (eg. texture-based, motion-based, tri-dimensional) retrieved from optical sensors can be complementary. Nevertheless, gathered measurements are susceptible to uncertainty, whether from noise introduced by the environment

or the sensor itself. Fusing information from distinct sources is a multidisciplinary solution to overcome the inherent weaknesses of each sensor/technique [39]. A single representation of the scenario that maintains the advantages of each modality can further reduce the impacts of sub-sea phenomena [38].

Perception systems are often integrated as payloads on mobile vessels for underwater interventions. In controlled environments, these optical setups have facilitated successful operations, such as valve turning [65], cable plugging and unplugging [66], and black-box recovery [67]. However, the majority of these works rely on human inputs, long surveys of the area and *a priori* knowledge to identify and retrieve the target objects. Complex algorithms are often employed to manage occlusions caused by the robotic arm, as poor visual feedback can significantly hinder intervention missions [68]. In eye-in-hand configurations, the perception system is mounted directly onto the robotic manipulator. While research is limited in applying this approach to underwater scenarios, it typically involves a single source of imaging data [37]. The additional degrees of freedom provided by the manipulator can be leveraged to inspect and monitor submerged structures closely. This enables the retrieval of information from hard-to-reach angles without the need to endanger the vessel by bringing it into close proximity with the target.

2.2 The Challenges of Underwater Perception

2.2.1 Sub-Sea Physical Phenomena

Vision-based perception systems face additional challenges when deployed in underwater environments. The propagation of light underwater is primarily influenced by two key physical phenomena: absorption and scattering. As a result, submerged optical sensors become limited not only in range but also in their ability to capture consistent and reliable data.

The absorption of light underwater can be defined as the loss of power through heat as the wave travels in the medium. Light attenuation is influenced by the medium's refractive index¹, the wavelength of the light, water salinity, and the concentration of dissolved particles [69]. Figure 2.1 presents the spectral radiance of light for multiple depth ratings, demonstrating that different wavelengths are absorbed at varying rates [70]. Longer wavelength colors, e.g. red, are mostly absorbed at 3 to 5 m depth. Contrarily, short wavelength colors are attenuated slower, leading to the typical blue color shift in underwater images. Additionally, this phenomenon significantly limits the effectiveness of Infrared (IR)-based technologies, such as time-of-flight sensors and structured IR light projectors (e.g., Light Detection and Ranging (LiDAR) or KinectV2), as IR wavelengths are highly attenuated within just 20 cm [71]. Natural illumination becomes insufficient beyond a certain depth. The use of artificial lighting is required for optical systems to operate in these conditions. Faulty choice of light sources and poor setup in relation to the optical system can lead to non-uniform illumination of the scene, causing vignetting, contrast and brightness inconsistencies [72]. Post-processing techniques have been explored in the literature for artificial light correction

¹Ratio between the speed of light in the vacuum and the speed of light in the medium.

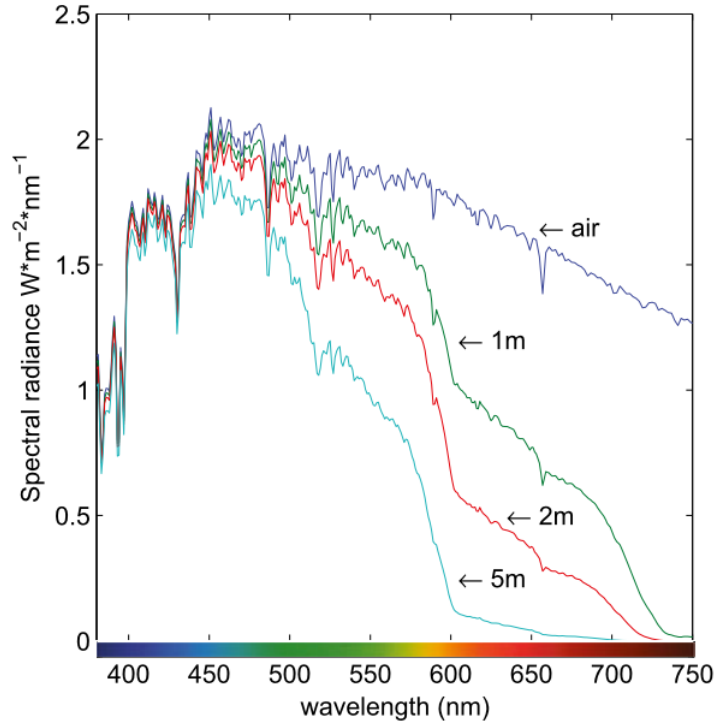


Figure 2.1: Simulation of a light wave traveling through water at different depths. Colors drop one by one, starting with the longest wavelengths. The absorption rate varies with water conditions, accelerating in high-salinity or turbid waters. Both suspended particles and dissolved organic matter further influence the absorption process. Extracted from [81].

[73, 74]. In turn, color correction of underwater images can be achieved through model-based approaches [75, 76], bilateral filters or adaptive histogram-based linear transformations [77, 78], and learning-based architectures [79, 80].

The existence of suspended particles in water, known as marine snow, severely degrades the visibility of submerged optical sensors. The aggregation of organic material and debris not only increases water turbidity but also causes light to deviate from its straight-line path, resulting in scattering [82]. This phenomena can be decomposed into two distinct components: forward scattering, where light reflected by the object is scattered into the optical system, causing image blurring; and backscattering, where light is reflected by suspended particles and reaches the camera, reducing contrast and limiting visibility [83]. Figure 2.2 shows an image affected by scattering phenomena, captured using an off-the-shelf imaging sensor during harsh environmental trials. Further analysis can be conducted if the Lambert-Beer empirical law is applied. Considering an isotropic and homogeneous medium, the total irradiance E at position r can be described as a function of both the absorption and scattering components of the physical medium, and given by equation 2.1 [83]:

$$E(r) = E(0)e^{-\alpha \cdot r}e^{-\beta \cdot r}, \quad (2.1)$$

where $E(0)$ represents the initial irradiance value, and α and β correspond to the absorption

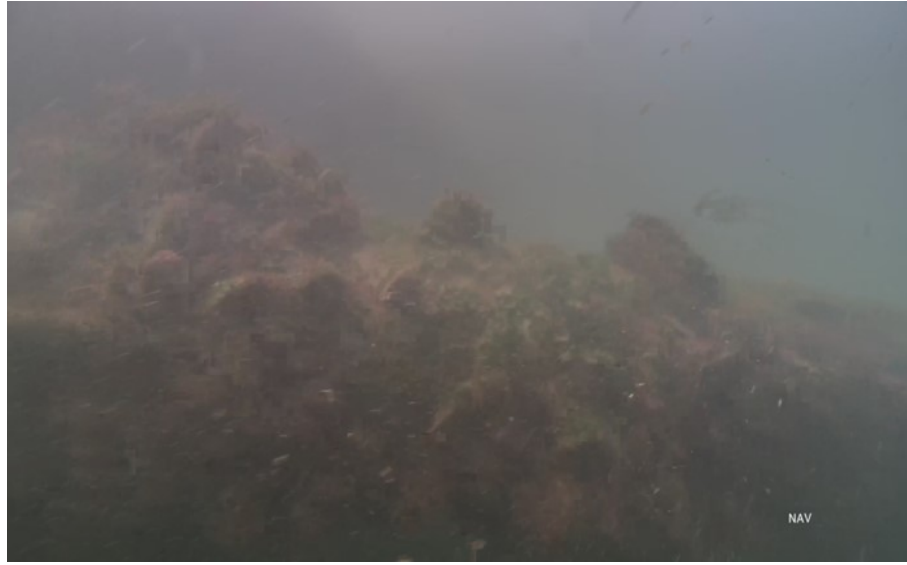


Figure 2.2: The impact of backscattering during image acquisition, as captured by an off-the-shelf imaging sensor.

and scattering coefficients, respectively. The scattering coefficient can be further defined as the sum of all scattering events. The probability function $B(\theta)$ expresses the likelihood of light being deviated by a certain angle θ . Finally, the coefficient β can be written as:

$$\beta = 2\pi \int_0^\pi B(\theta) \sin(\theta) d\theta. \quad (2.2)$$

According to the Jaffe-McGlamery image formation model [84, 85], in underwater scenarios, the total light received by an optical system (E_T) can be expressed through equation 2.3:

$$E_T = E_d + E_b + E_f, \quad (2.3)$$

where E_d is the direct component that represents the light reflected by the target scene; and E_b and E_f represent the backwards and forwards scattering components. The scattering phenomena can be mitigated directly during the image formation process through range-gated imaging techniques [86, 87, 88]. This concept consists in adjusting the camera's shutter period so that the sensor only receives light reflected from the scene, and not the scattered photons, increasing robustness to both the E_b and E_f components [82]. While longer exposure times allow for larger depth ranges to be acquired, these will also increase the susceptibility to the effects of backscattering. The captured underwater images can be post-processed for increased fidelity through filtering and contrast equalization methods [89], dark channel priors [90, 91], model-based restoration algorithms [69, 92], and learning-based solutions [93, 94].

Visual context for the discussion on sub-sea physical phenomena is provided in figure 2.3. Absorption and scattering present significant obstacles to underwater perception systems. These challenges are heightened by the presence of suspended particles, which hinder the acquisition of high-quality data and drastically limit the effective operation range.

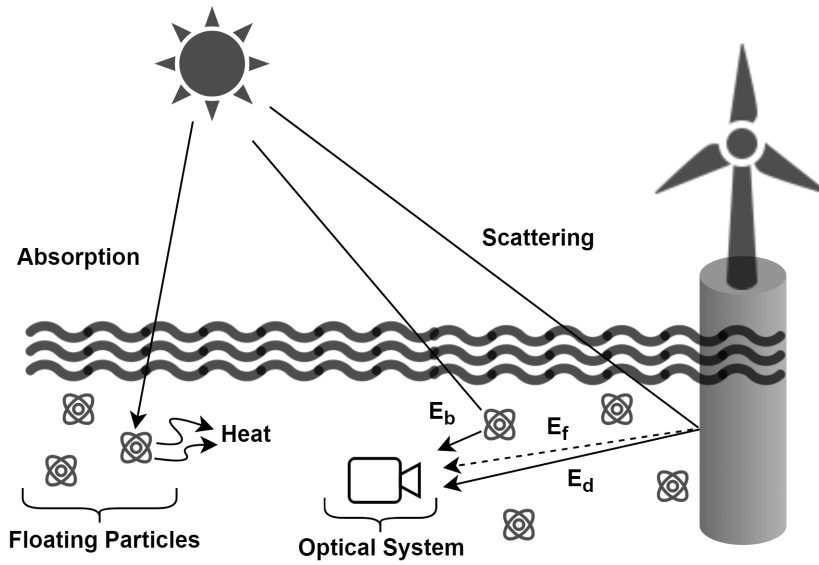


Figure 2.3: A visual representation of the two main sub-sea phenomena affecting underwater optical systems: absorption, where the energy of light is dissipated as heat when interacting with molecules in the medium, and scattering, where light deviates from the target scene and is scattered onto the sensor.

2.2.2 Calibration of Underwater Optical Systems

The successive refraction of light through water, glass, and air introduces optical distortions that must be addressed to ensure accurate measurements [95]. The mechanical housing that protects and waterproofs the imaging sensor significantly affects the quality of captured underwater images, as the geometry of the viewport and its alignment with the sensor impact how light behaves when passing through these layers [96]. Figure 2.4 demonstrates the refraction effects induced by both flat and hemispherical dome ports.

The refraction of light at flat interfaces causes rays reflected from single object to appear as though captured from multiple viewpoints - refer to figure 2.4a. These apparent viewpoints stem from the lack of collinearity between the target point in the scene, the center of projection of the optical system and the image plane point [97]. Flat viewports have reduced Field-Of-View (FOV) and the induced radial distortion increases with respect to the distance from the center of projection [98]. Additionally, optical systems in this configuration that assume a single viewpoint model, estimate measurements with significant inaccuracies [97]. Hemispherical interfaces effectively reduce light refraction effects [96], as illustrated in figure 2.4b. By aligning the optical system's center of projection with that of the dome interface, light rays that pass through the center of projection are orthogonal to both the inner and outer surfaces of the dome, minimizing distortions [99]. These interfaces are generally preferred for underwater optical systems, as a careful alignment enables the use of state-of-the-art camera calibration techniques. Nevertheless, even if the optical system is pre-calibrated on air, it should be fine-tuned for underwater scenarios, as variations in water temperature, pressure, and salinity can directly affect calibration parameters [100].

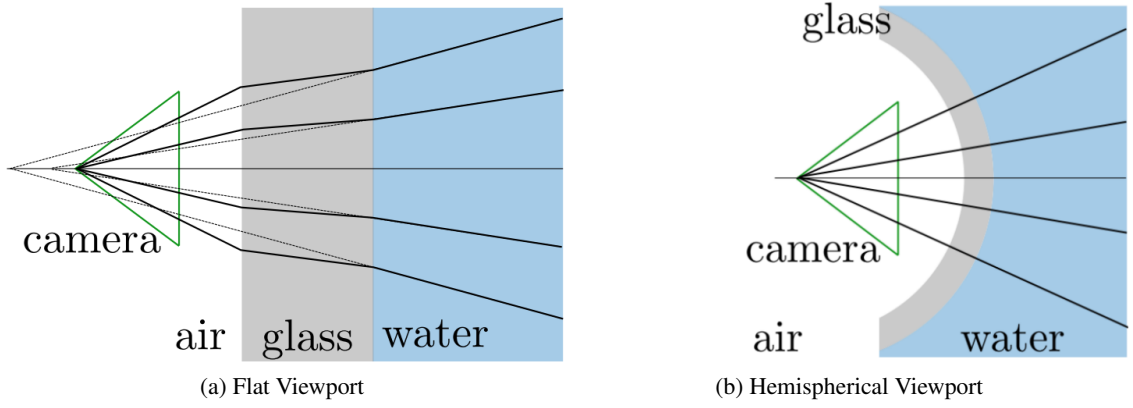


Figure 2.4: The refraction of light caused by the interactions between two mediums, for distinct viewport geometries. Extracted from [96].

Most conventional calibration methods rely on the pinhole camera model for underwater systems due to its straightforward mathematical structure and ease of implementation. However, this model does not consider the multilayer refraction of light, which can introduce substantial uncertainties in 3D measurements unless distortion models are incorporated [101]. The adaptive pinhole model addresses this limitation by accounting for light interactions across multiple media. For a camera setup with its optical axis approximately perpendicular to the interface (whether planar or hemispherical), refraction effects can be effectively compensated by adjusting focal length and distortion parameters [102, 103]. Adaptive calibration techniques often involve using patterns submerged at various distances and angles, though this approach limits the measurements outside the calibration range [104]. Recent adaptive approaches use learning-based methods to adjust distortion parameters; these methods, however, still lack the precision required for highly accurate underwater reconstructions [105, 106].

Radial distortion correction methods assume that primary refraction effects are radially symmetric around the principal point, for optical setups where the camera's axis is approximately perpendicular to the housing interface [107]. Nonlinear lens distortions cause radial displacement of image points: barrel distortion curves straight lines outward from the center, while pincushion distortion curves lines inward. This radial displacement can be modeled using polynomial functions to mathematically represent and correct these distortions [108]:

$$r_d = r_u \cdot (1 + k_1 r_u^2 + k_2 r_u^4 + k_3 r_u^6 + \dots), \quad (2.4)$$

where r_d is the distorted radius, defined as the distance from the image center to a point in the distorted image; r_u is the undistorted (ideal) radius, defined as the distance from the image center to a point in the undistorted image; and k_1, k_2, k_3 are distortion coefficients obtained through camera calibration. The radial component of the calibration parameters can be adjusted to compensate for this refraction effect [109, 110]. For measurements outside the calibration range, the pinhole camera model becomes inappropriate [111]. Additionally, the maximum reconstruction errors

tend to appear at the periphery of radially corrected images [97].

Ray tracing geometric correction methods simulate the precise path of light rays to understand their interactions with various surfaces and materials, enhancing the accuracy of correcting lens distortions [112, 113]. A. Agrawal *et al.* (2012) [114] demonstrate that the geometry of light rays in flat refraction systems aligns with an axial camera model, where the optical axis is perpendicular to the housing interface, resulting in symmetric light reflections. Similarly, X. Chen and Y. Yang (2014) [115] extended this geometric refractive calibration method to stereo configurations by estimating the interface thickness directly from stereo pixel correspondences. For dome interfaces, M. She *et al.* (2022) [116] show that refractions also align with the axial model, as light intersects the hemispherical dome at two poles, leading to barrel or pincushion refraction effects. A non-iterative calibration procedure is introduced to compute the center of refraction without prior knowledge of the media's properties. Other geometric approaches leverage wavelength dispersion, exploiting how different wavelengths of light project onto distinct image points due to varying refractive indices [117], or estimate the pose of the transparent interface to obtain initial values for camera intrinsic parameters, which are later refined using a quaternion-based approach [118]. Alternatively, M. Roznere *et al.* (2024) [119] propose an N-Sphere camera model where points from the image plane are mapped onto an N-dimensional sphere, efficiently handling the distortions typical of wide FOV and fisheye lenses. The model treats each medium transition as additional spheres in the optical path, introducing parameters to handle camera offsets within the dome while maintaining computational efficiency.

2.3 Underwater Tri-dimensional Scene Understanding

2.3.1 Reconstruction Techniques and Imaging Sensors

A myriad of solutions have been explored to provide autonomous underwater systems 3D information about their surroundings. Vision-based sensors can retrieve dense and textured information in close proximity. Imaging techniques can be arranged into passive or active approaches, providing data with different characteristics [120]. Learning-based solutions extract features directly from data to obtain robust depth estimation models [121]. Domain-specific adaptations can mitigate the limited amount of available information.

Passive Imaging

Passive underwater imaging makes use of natural or artificial light sources to capture tri-dimensional information from the scene. Photogrammetry-based techniques estimate pixel-to-pixel displacements (disparity) between images taken from distinct viewpoints. By knowing the fixed relative positioning and orientation of the cameras, it is possible to triangulate 3D points, as depth is inversely proportional to the disparity. The 3D scene representations obtained from these methods offer high spatial resolution and dense coverage in well-textured areas, with color information mapped to each point. However, water turbidity and suspended particles introduce noise



Figure 2.5: The cylinder-shaped AQUASENSOR houses an off-the-shelf stereo camera used by divers for reconstructing and mapping corals and shipwrecks. Extracted from [125].

and false pixel correspondences, leading to spurious points. Additionally, regions with uniform texture often yield unreliable reconstructions due to ambiguity in correspondence matching [122].

The Invariant 3D Reconstruction by Instrumented Stereovision (IRIS) is an adjustable, stereo-based system introduced by A. G. Allais *et al.* (2005) [123]. Designed for deep-sea applications, each camera is depth-rated for 6000 m, with one camera mounted on a pan-tilt mechanism that allows for adjustable angles and distances between the cameras. IRIS was integrated as payload on the Victor6000 ROV to conduct 3D reconstructions and texture mapping in underwater environments [124]. The AQUASENSOR uses a Bumblebee² stereo camera housed in a robust, cylinder-shaped enclosure rated for depths up to 30 m [125] - refer to figure 2.5a. Developed by A. Hogue *et al.* (2007) [125], this sensor combines stereo vision with Inertial Measurement Unit (IMU) data to estimate 3D scene representations at 2-4 Hz. AQUASENSOR has been handheld by drivers in both shipwreck and coral reef reconstruction missions (figure 2.5b). P. Carrasco *et al.* (2015) [126] also use an off-the-shelf Bumblebee stereo camera mounted on the SPARUS II AUV, with the optical axis positioned perpendicular to the ground. Simulation results show that the proposed Stereo Graph-SLAM algorithm significantly improves the vehicle's localization, reducing the mean trajectory error by up to 80%. The work of K. Williams *et al.* (2014) [127] focus on the study of biological activity through a simple off-the-shelf stereo camera pair encapsulated with a hemispherical interface, capable of reaching 500 m depths. The TrigCam system illuminates scenes with far-red light and captures high-resolution images when fauna enter the camera's field of view, allowing for non-intrusive observation of marine life. A modular stereo setup comprised of two Allied Vision Technologies (AVT) G125C³ cameras is introduced by F. Oleari *et al.* (2015) [128] to aid in underwater grasping and manipulation tasks. The system provides high-resolution disparity maps (1292x964) at a rate of 12.5 Hz.

²<https://www.flir.com/support/products/bumblebee2-firewire>

³<https://www.1stvision.com/cameras/models/Allied-Vision/Mako%20G-125B/C>

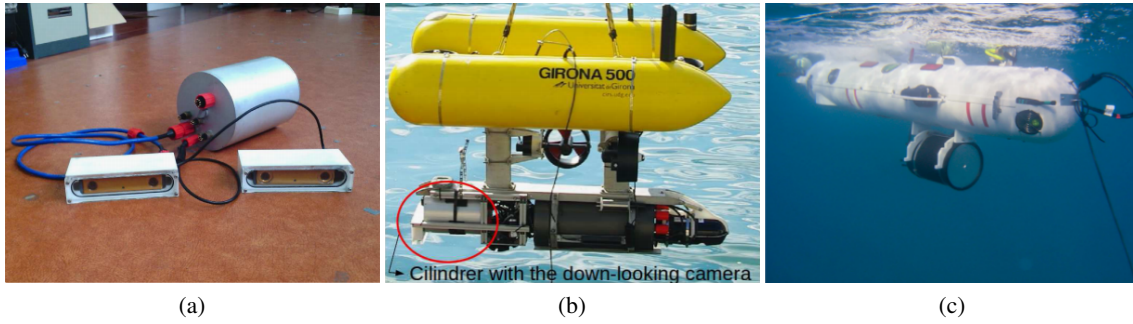


Figure 2.6: The Fugu-f optical system is designed to be a flexible sensor capable of retrieving high-resolution tri-dimensional information. It has been equipped as payload in vehicles such as the Girona500 - figure b); and Nessie VI - figure c). Extracted from [61].

The research of F. Bonin-Font *et al.* (2015) [61] presents the Fugu-f modular imaging sensor, developed to enhance the intervention capabilities of underwater mobile robots. The system includes two Bumblebee stereo cameras housed in flat-ported enclosures, with power and processing units in a separate cylindrical module. Fugu-f has demonstrated versatility in tasks such as mosaicking and object detection, achieving a maximum depth resolution of 0.0107 m at 1 m distance. Figure 2.6 shows the sensor integrated as payload in various underwater vehicles. M. Johnson-Roberson *et al.* (2017) [129] introduce the Diver-Rig, a stereo setup that pairs auto-exposed 1360×1024 AVT Prosilica Red, Green, Blue (RGB) and grayscale images at a rate of 2 Hz. The system is designed for shallow water mapping, ensuring a spatial resolution 1 mm/pixel. The methodology developed by G. Huo *et al.* (2018) [130] employs stereo matching principles in restored underwater images. An absolute error of 0.0085 m is obtained for the calibration distance of 1 m; as this distance increases, the mean relative error reaches 5.26% (0.11 m). A stereo vision-based ROV docking approach is developed by M. Myint *et al.* (2018) [131]. Controlled environment experiments validate this approach, as the low resolution camera pair (640 × 480) guides the vehicle towards a known 3D marker with less than 5 mm pose error. The PelagiCam module, introduced by E. V. Sheehan *et al.* (2020) [132], is designed to monitor marine fauna with minimal disruption. Each module consists of three GoPro HD Hero2 cameras positioned 120° apart to capture a full 360-degree view of the surrounding environment. This imaging setup has proven effective for detecting and recording marine fauna in a non-destructive, non-intrusive manner. Y. Ma *et al.* (2023) [133] employ photogrammetry principles for detecting cracks and measuring widths in underwater structures. This setup incorporates a pair of commercial cameras (1200×960 resolution) housed in a flat-ported cylindrical enclosure. In controlled testing, this system achieved a mean relative assessment error of 7.8%. A two-stage stereo depth map estimation approach is presented by W. Wang *et al.* (2023) [134]. The proposed pipeline closely matches the state-of-the-art COLMAP [135], achieving accuracy and completeness within 0.05 m. S. Tani *et al.* (2023) [136] employ a stereoscopic system for AUV navigation, using downward-facing, high-resolution cameras (1920×1440) triggered at 2 Hz to support a visual odometry algorithm. Under optimal visibility conditions, validation tests reveal trajectory errors of 7% and 13% across sepa-

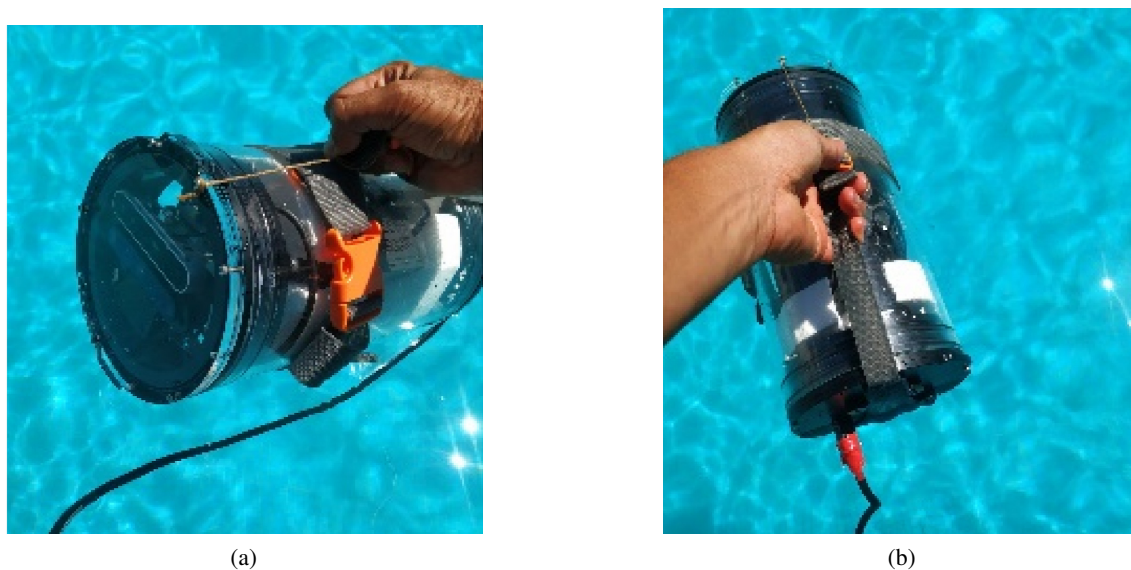


Figure 2.7: A low-cost stereo vision system developed from off-the-shelf technology for fish length measurement. The optical configuration ensures a 10 cm baseline and 60° FOV. Extracted from [141].

rate survey missions. The spatial dimensions of abalone shells are measured using the binocular system proposed by C. Yuehang *et al.* (2024) [137]. Using a Pyramid Stereo Matching Network (PSMNet) [138] to obtain disparity values, the system achieves an average measurement error of 2.02% for length and 2.54% for width. The Lanty stereo vision system was designed by E. Alfaro-Dufour *et al.* (2024) [139] for fish monitoring. It features two 2048x1536 resolution Flir Chameleon⁴ cameras housed in an acetal copolymer and stainless steel structure rated to 500 m depth. The system uses the state-of-the-art Semi-Global Block Matching (SGBM) [140] for disparity estimation, while a You-Only-Look-Once (YOLO) v8 algorithm is used for detection and classification. Real-world field tests were successfully conducted along the coasts of Mallorca. C. Silva *et al.* (2024) [141] also target the aquaculture sector, developing a stereo system for fish length measurement. This imaging sensor, illustrated in figure 2.7, uses an optical configuration with 10 cm baseline and 60° FOV, integrated within a watertight flat ported acrylic chamber. Both simulations and real-world tests validate this setup, demonstrating its capability to measure fish lengths with less than 1% error.

Image sequences captured from monocular optical setups can be used to infer 3D representations of a scene. Photogrammetry principles can be applied if the relative camera motion between consecutive frames is known [142]. Structure-from-Motion (SfM) algorithms provide a dense, textured point cloud; however, SfM cannot determine the absolute scale of the reconstructed scene without accurate navigation data to track the camera’s position and orientation. Additionally, it may struggle with non-textured surfaces or dynamic scenes [143, 144]. The work of A. Sedlazeck *et al.* (2009) [145] adapts SfM techniques for underwater environments by filtering out back-

⁴<https://www.teledynevisionsolutions.com/products/chameleon3-usb3/>

ground noise and floating particles. Color correction methods are employed to improve robustness in camera pose estimation and enable accurate reconstruction from a sparse set of points. C. Beall *et al.* (2010) [146] expand this concept to a 1920×1080 resolution stereo rig with 60 cm baseline. An initial camera pose is obtained from feature detection and matching, which is later refined by a smoothing-based SLAM algorithm. An SfM reconstruction of coral reefs is discussed in the research of D. Mckinnon *et al.* (2011) [147], using high-resolution images (2272×1704) to generate a detailed 3D representation. The reconstruction achieves errors of less than 1 mm for 90% of the coral surface, validated against a laser scan ground truth. A. Jordt-Sedlazeck and R. Koch (2013) [148] present an SfM method that accounts for light refraction at flat port interfaces during camera pose estimation. Compared to traditional SfM, this approach reduces re-projection errors in both synthetic and real underwater reconstructions, resulting in improved accuracy. Ray tracing of light refracted from flat planes is explored in the research of A. Shibata *et al.* (2016) [149] to address the scale ambiguity in SfM without requiring additional navigation data. Simulation experiments reveal that light rays can be accurately traced to adjust the scale of the resulting 3D representation, if the thickness of the plate and its distance from the camera are known *a priori*. X. Xu *et al.* (2016) [150] estimate tri-dimensional SfM representations using low-resolution cameras. Images captured in challenging underwater conditions with ambient lighting of 0.28 lx are processed to filter noise. Features are then extracted using the Scale Invariant Feature Transform (SIFT) algorithm, tracked, and combined through SfM and patch-based multi-view stereo techniques. A new Refractive Structure-from-Motion (RSfM) pipeline is introduced by F. Chadebecq *et al.* (2017) [151], which enhances initial camera pose estimation through a refractive fundamental matrix derived from epipolar constraints. In underwater tests using an off-the-shelf GoPro camera, the RSfM algorithm achieves reconstruction accuracy with a root mean squared error of 4.6 mm and a standard deviation of 3.2 mm. This research is further extended into a two-view SfM algorithm [152], introducing a novel formulation that incorporates an rF constraint, enabling robust quasi-dense reconstruction in underwater environments.

K. Istenic *et al.* (2020) [153] tackle the scale ambiguity problem in SfM by inferring scale from a configuration of laser beams projected onto the environment (see figure 2.8a). The authors propose an unconstrained for arbitrary laser setups and a partially constrained method that assumes parallel lasers at a fixed camera distance. The latter method shows an overall lower performance (2.9% for the worst case scenario) due to the approximated laser direction. Real scenario experiments employ the fully unconstrained method to estimate that each unit of the 3D model corresponds to 0.237 m - refer to figure 2.8b. An efficient global SfM method is developed by Y. Chen *et al.* (2022) [154], where the relative camera pose estimation step is refined using the Least Trimmed Square Rotation Averaging (LTS-RA) approach. Experiments in controlled scenarios show that this method achieves approximately 70% lower transmission costs without any loss in metric performance. M. She *et al.* (2024) [155] improve upon the state-of-the-art COLMAP algorithm to support generalized refractive camera setups, accommodating both flat and dome-port housings. Additionally, a robust relative pose estimation approach ensures reliable SfM initialization. Evaluations in real-world scenarios demonstrate the effectiveness of this approach, with the

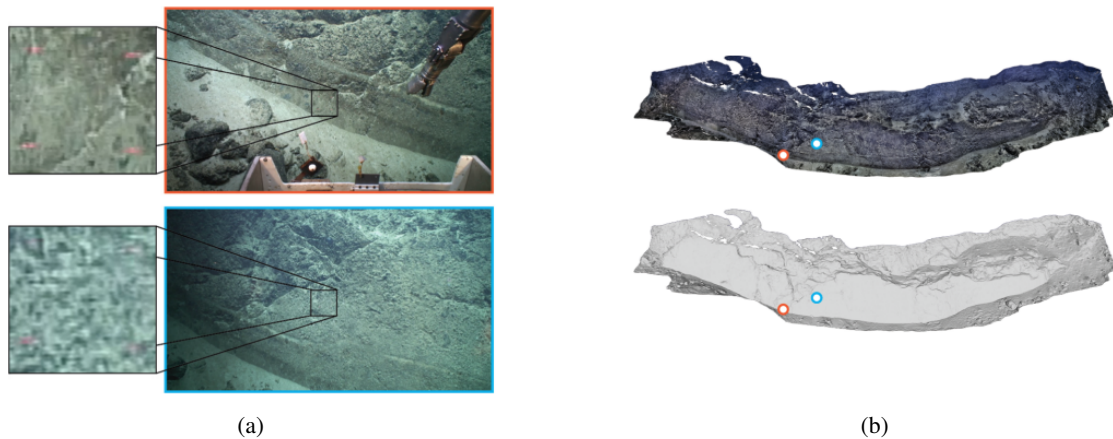


Figure 2.8: Laser beam projections on the scene with known patterns that allow for the scale of the SfM reconstruction to be estimated - figure a). Final tri-dimensional reconstruction is depicted in figure b). Extracted from [153].

RSfM pipeline achieving reconstruction errors as low as 2.103 mm.

Active Imaging

Underwater active imaging employs the projection of laser stripes, dots, or structured light patterns onto underwater scenes. Tri-dimensional information is derived by segmenting and analyzing the deformation of these patterns as captured by a camera. This structured light approach is capable of producing high-resolution, accurate 3D representations, which are particularly valuable for close-range applications. However, underwater light scattering and absorption can cause rapid degradation in pattern quality over distance, limiting effective range. Moreover, these systems demand precise calibration to account for optical distortions and typically require higher power, presenting challenges for long-duration operations [101].

Single-laser structured light configurations rely on the precise positioning between the laser source and the camera, using triangulation to retrieve spatial information. H. Kondo *et al.* (2004) [156] proposed a system that uses this approach to navigate vehicles by tracking artificial underwater structures, achieving millimeter-level accuracy within a 2 m. Similarly, A. Yamashita *et al.* (2004) [157] refined this approach by integrating a light refraction model, which improved close-range accuracy to within 1 mm. The study presented by G. Inglis *et al.* (2012) [158] applies this principle to obtain a high-resolution bathymetry. The information extracted from the laser stripe is stitched together using a SLAM algorithm, resulting in a grid map with accuracy up to a centimeter. A. Bodenmann *et al.* (2013) [159] integrate a single-laser imaging system on the Hyper-Dolphin ROV to a seafloor reconstruction with average pixel resolution of 6.4 mm. M. Massot-Campos and G. Oliver-Codina (2014) [160] explore the use of diffractive elements to expand the range of information gathered by the single laser-camera setup - refer to figure 2.9. This strategy allows for a more complete 3D reconstruction but its effectiveness quickly decreases for higher water turbidity levels, as the diffracted light is attenuated faster. N. Hansen *et*

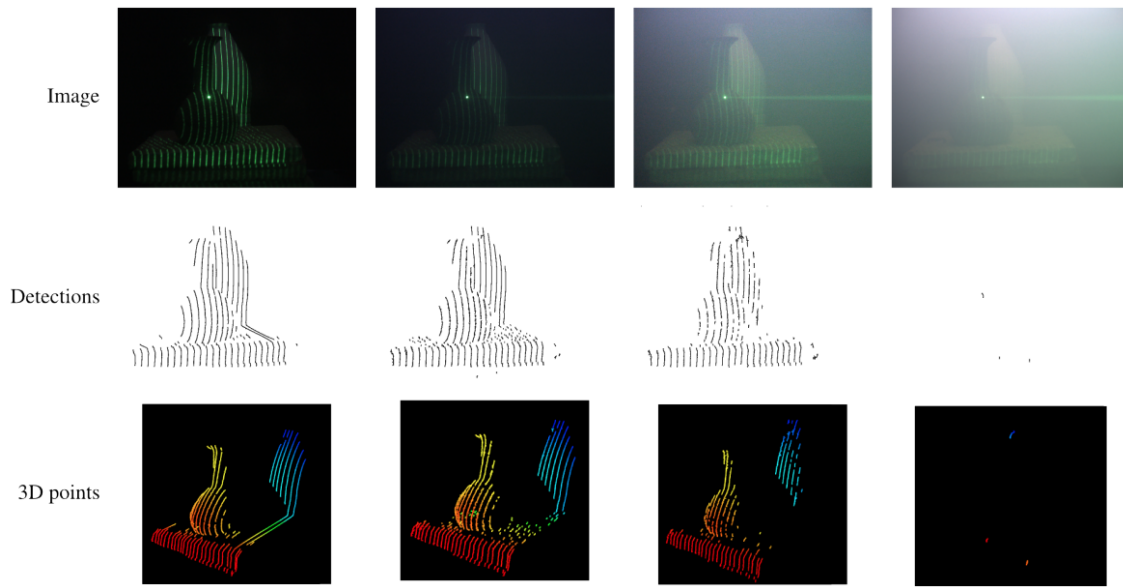


Figure 2.9: A single laser beam is diffracted into multiple lines, with lower-powered diffracted lines attenuating first as turbidity levels rise. Extracted from [160].

al. (2015) [161] integrate a two-laser configuration within an OpenROV vehicle to perform underwater inspection tasks. The system was tested both in simulation and experimentally, achieving measurements with an approximate error of 0.04 m. An array of three line-lasers coupled with a camera is used by C. Constantinou *et al.* (2016) [162] to evaluate the relative positioning between an ROV and aquaculture fish-net targets. Testing in controlled laboratory environments demonstrate measurement errors within 2.5%. A fixed camera and a hand-held cross-line laser projector configuration is used by M. Bleier and A. Nüchter (2017) [163], leveraging co-planarity geometric constraints to reconstruct depth information. G. Matos *et al.* (2020) [164] improve laser calibration accuracy by accounting for camera distortion and light refraction. Simplistic underwater measurements reveal a pixel resolution of 0.01 mm. S. Li *et al.* (2023) [165] divert from the classic laser line threshold segmentation process by employing a state-of-the-art semantic segmentation Pyramid Scene Parsing Network (PSPNet) [166]. A normalized grayscale gravity method is then used to accurately pinpoint the stripe's centerline. Tests in a water tank demonstrate that this approach is both robust to noise and generalizes effectively. A novel underwater pipeline tracking method is proposed by J. Fan *et al.* (2024) [167] using a dual-line laser active imaging setup. The laser stripe-based 3D reconstructions are combined with Doppler Velocity Log (DVL) data to precisely calculate relative positioning to stationary pipelines. The optical setup is integrated in the BlueROV vehicle and controlled underwater experiments show pipeline diameter measurement errors down to 0.0045 m. A multicolored light array is used by S. Li *et al.* (2024) [168] to obtain a broader one-shot perception of the environment. Laser segmentation is conducted through a HSV double-threshold. Experimental trials show a close-range tri-dimensional accuracy of 0.63 mm.

Structured light perception systems can also make use of pattern projectors to obtain dense point clouds through triangulation, or to creating artificial texture that benefits block matching

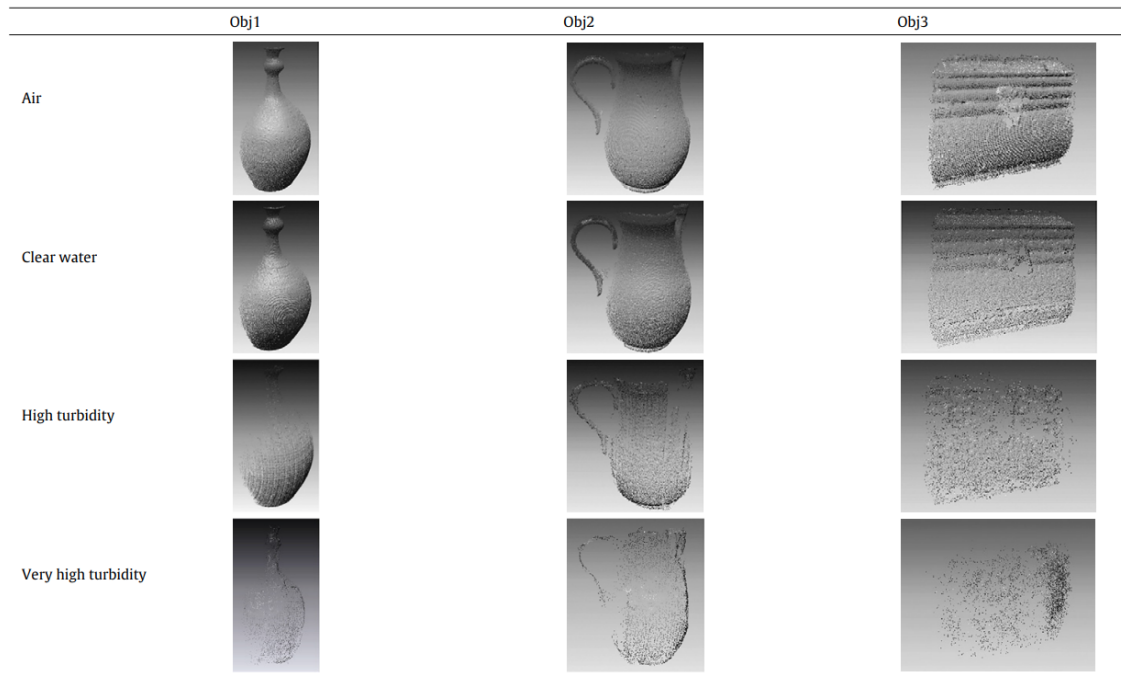


Figure 2.10: Gray-coded pattern projector improves the matching of a stereo composition. As turbidity levels increase, the quality of the estimated reconstructions quickly degrades. Extracted from [170].

methods. A variety of patterns have been used in the literature, e.g. gray-code bands, sinusoidal fringes or grids [169]. These provide broader spatial coverage of the environment at the expense of increased power consumption. Harsh underwater environments can also vastly degrade the performance of pattern projectors as these are quickly absorbed in turbid waters [170, 171]. A stereo optical setup combined with a gray-coded pattern projector is used by F. Bruno *et al.* (2011) [170] to improve the quality of the estimated disparity map. Figure 2.10 illustrates the 3D point clouds obtained in underwater scenarios with varying turbidity levels. Compared to the baseline measurements taken in air, underwater reconstruction quality degrades as turbidity increases, with the number of points per 100 pixels reducing from 13.42 to 1.38. Additionally, a long acquisition time renders the system unusable for real-time applications. Q. Zhang *et al.* (2011) [172] introduce a method for estimating the 3D shape of submerged objects using phase tracking and ray tracing algorithms. A digital projector illuminates the object with sinusoidal fringes, and the interactions with the object's surface are modulated. Homologous points in the fringes are identified through phase tracking, and each corresponding ray is then traced to determine the 3D coordinates of these points. This method is reportedly faster than a conventional stereo-based approach, at the cost of lower accuracy in object reconstruction. A customized grid-based pattern laser is used in combination with a camera in the work of H. Morinaga *et al.* (2015) [173]. The projected grid pattern aids a polynomial approximation method for estimating the system's distortion model. To enhance accuracy, the initial 3D reconstructions are refined using a bundle adjustment algorithm, which optimizes the reconstruction by reducing projection errors. A. Sarafraz and B. Haus (2016)

[62] employ an underwater camera setup with an external pattern projected from above the water surface. The reconstruction approach accounts for the distortions introduced by both the object's surface and the water interface. While this method achieves precise reconstructions, it relies heavily on water clarity, limiting its applicability for real world scenarios. Y. Nishida *et al.* (2021) [174] project a De Bruijn-coded laser pattern [175], using six colors spaced 60° apart on the Hue scale. The decoding process relies on a predefined table, achieving relative measurement errors of less than 1.4%. Speckle patterns are studied in the research of S. Zhuang *et al.* (2023) [176] to improve the accuracy of stereo-based techniques. A new row entropy parameter is introduced to evaluate the influence of the added texture on the tri-dimensional representations. Real underwater experiments show that a small row entropy coefficient leads to more accurate reconstructions.

Active imaging trends towards rotating laser-line scanners, which allow a single beam configuration to obtain a dense 3D representation at the expense of a more complex to assemble module. F. Lopes *et al.* (2015) [177] introduce a system composed by a digital camera and a rotating laser projector, capable of reconstructing an object with a global error of 2% of its size. The work of S. Chi *et al.* (2016) [178] uses a galvanometer to adjust the direction of a laser stripe, enabling automatic scanning for 3D reconstructions of underwater scenes. Experimental results under ideal underwater conditions reveal a maximum distance error of 1.87 mm. A 3D laser scanner that also makes use of a galvanometer to adjust the angle of a mirror is proposed by A. Palomer *et al.* (2018) [47, 179] (see figure 2.11). The system operates at up to 6 Hz with a maximum angular resolution of 0.008° between projected lines. Under optimal lighting (e.g., a dark environment for clear line segmentation and triangulation), it achieves an average error of 0.53 mm with a corresponding standard deviation of 0.27 mm. This active imaging system was later integrated into the Girona500 AUV to support grasping and manipulation tasks. A modular single camera-laser configuration is presented by Q. Xue *et al.* (2021) [180]. The system is mounted on a turntable that defines the operating angle. Simple trials reveal a reconstruction error of 0.6 mm within the working range of 0.14 to 2.5 m. C. Gu *et al.* (2021) [181] introduce a unified laser-based 3D scanner calibration method that operates across different media without recalibration. Additionally, an underwater 3D scanner prototype is developed, demonstrating sub-millimeter precision reconstructions of various objects in multi-medium scenarios. A structured light 3D reconstruction system based on a

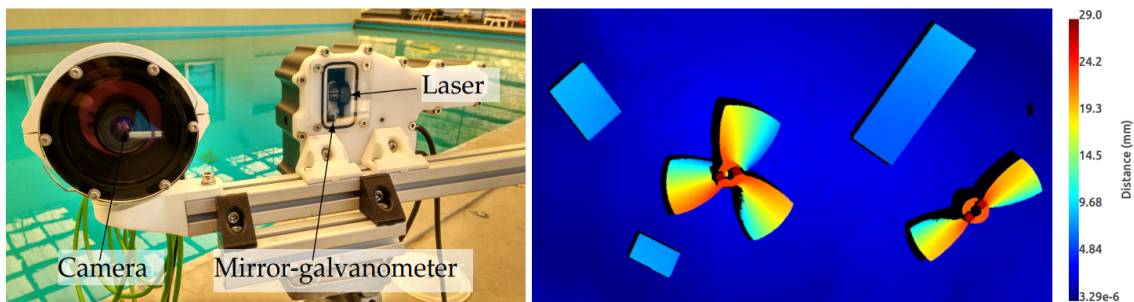


Figure 2.11: A galvanometer iteratively redirects the projected laser beam allowing for increased spatial coverage of the scenario. Extracted from [47].

rotating laser scanner is developed by Y. Ou *et al.* (2023) [182] that employs a binocular polar curve constraint model to ensure accurate line matching. This solution achieves a resolution of 1.025 mm in clear water at a distance of 0.5 m. Experiments in turbid waters show metric errors smaller than 3 mm. S. Li *et al.* (2023) [183] combine line structured light with an optical oscillator to assemble a monocular self-scanning measurement system. The rapid reflection of the laser beam allows for a complete sweep of the target, while a refraction compensation method enhances the final point cloud. In controlled trials, the system achieves 0.398 mm accuracy, reconstructing submerged objects with relative errors below 2%. A cost-efficient structured light system based on a sliding scanner is introduced by H. Lin *et al.* (2024) [184], using ray tracing to establish a refraction-based imaging model. Separate calibrations in air and underwater refine the FOV transformation, correcting 2D image distortion and restoring accurate 3D point depth. An enhanced laser stripe extraction algorithm improves pixel accuracy, enabling reliable 3D reconstructions. Water tank tests reveal an error rate of approximately 2.89% when reconstructing standard concrete objects underwater.

Hybrid Imaging

Hybrid underwater imaging systems combine both active and passive techniques, integrating the benefits of each method to capture complementary information that enhances 3D reconstruction accuracy and scene understanding. However, hybrid systems are complex to design and assemble, requiring precise calibration and synchronized control units to manage both the timing and alignment of active and passive components. The intricate coordination and setup required make hybrid systems more suitable for specialized applications where high-resolution depth data is essential. The MARESys hybrid imaging sensor is first introduced by A. Pinto and A. Matos (2020) [87]. Information is retrieved from the scene via both photogrammetric stereo techniques and laser line triangulation, capturing dense texture-based and sparse accurate point clouds. A range-gated acquisition scheme guides the imaging formation of both 1280x960 AVT MAKO 125C cameras, a

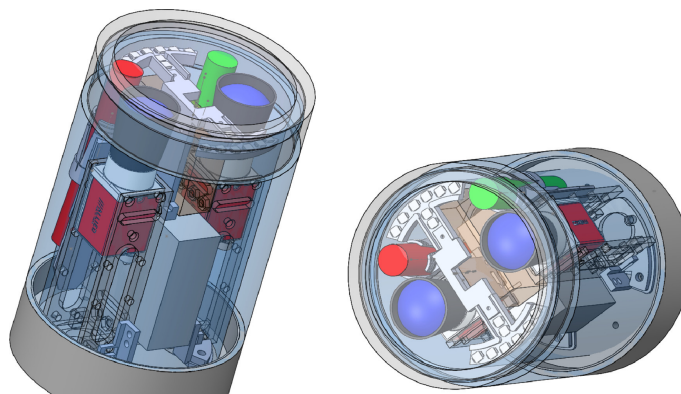


Figure 2.12: Concept design of the MARESys hybrid imaging sensor: this system integrates both passive and active imaging techniques for underwater applications. The passive stereo setup includes two AVT MAKO 125C cameras paired with LED lights, while active imaging uses two laser beams in combination with the cameras. Extracted from [87].

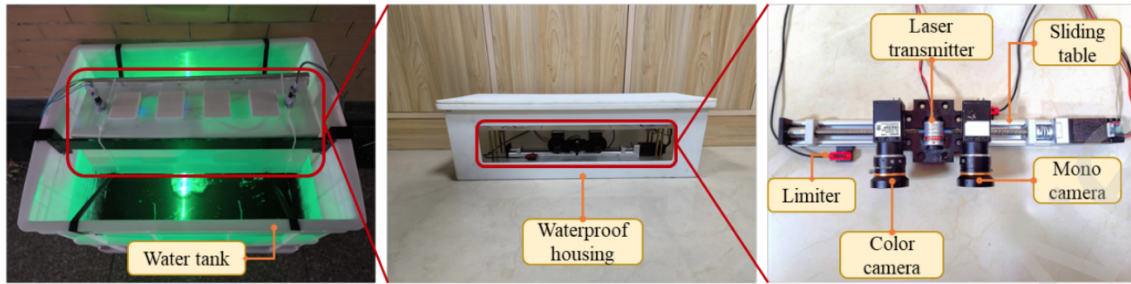


Figure 2.13: Hybrid imaging system composed of a stereo camera pair and a laser diode mounted on a sliding platform. Extracted from [185].

mitigating the impact of sub-sea phenomena to gather high fidelity data. In controlled experiments simulating different levels of water turbidity, the stereo-based method achieved an average error of 0.025 m, while the laser-based method demonstrated superior accuracy with an average absolute error of 0.010 m. H. Lin *et al.* (2024) [185] also propose a hybrid imaging system that includes a pair of 1440x1080 HIKROBOT⁵ cameras, a laser beam emitter, and a sliding platform to enhance spatial coverage. Figure 2.13 illustrates this setup. The system leverages mathematical models for underwater laser triangulation and enhanced stereo pixel matching. Water tank experiments demonstrate 3D reconstruction accuracy with relative errors under 5%. Although this system employs hybrid imaging, it does not fully take advantage of the principle, as the cameras are used primarily for calibration and point cloud texturing rather than as an additional source of 3D data.

Learning-based Solutions

Depth estimation from a single image is an ill-posed mathematical problem, as the same 2D projection can correspond to multiple 3D scene configurations. Recent advances in deep learning have produced powerful architectures capable of estimating dense point clouds with state-of-the-art precision, particularly in aerial applications [186, 187]. Underwater monocular depth estimation research stems from image enhancement tasks, which aim to improve visual clarity in underwater images [90, 188, 189]. On the other hand, learning-based models require extensive training data and often struggle to generalize across different underwater environments. The scarcity of publicly available datasets for underwater depth estimation further limits the development of AI networks for this domain [44]. H. Gupta and K. Mitra (2019) [190] propose an unsupervised network for depth estimation by adapting aerial hazy images to underwater contexts. This style-transfer model of learning employs two circularly connected dense-block auto-encoders, optimized through similarity and gradient sparsity losses. Experiments with the small Stereo Quantitative Underwater Image Dataset (SQUID) RGB dataset introduced by D. Berman *et al.* [191] demonstrate a 40.7% improvement in scale-invariant mean squared error over prior state-of-the-art models. A similar strategy is adopted by X. Ye *et al.* (2020) [192] where a Style Adaptation Network (SAN) is used to adapt in-air images to the underwater domain, enabling joint learning for depth estimation and color correction by focusing on domain-invariant features.

⁵<https://www.hikrobotics.com/en/machinevision/productdetail?id=4994>

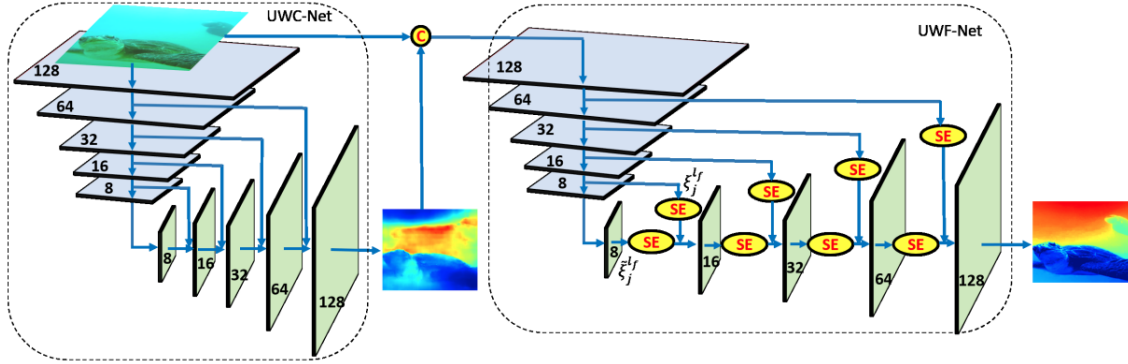


Figure 2.14: The UW-GAN is composed of two sub networks: UWC-Net estimates a coarse depth map; and UWF-Net refines this representation by combining it with the initial RGB input image. Extracted from [194].

However, discordant colors with the underwater tone can negatively impact the accuracy of the generated 3D representations. The work of J. Pérez *et al.* (2020) [193] estimates a coarse depth map with a neural network, refining it via a guided filter. This refined depth map is then used to de-haze sub-sea images through an inverse underwater image formation model.

The Underwater Generative Adversarial Network (UW-GAN) introduced by P. Hambarde *et al.* (2021) [194] is illustrated in Figure 2.14. In the first stage, the UWC-Net estimates a coarse depth map, which is then refined by UWF-Net, a fine-level network utilizing spatial and channel-wise squeeze-and-excitation blocks to combine the initial depth map with the input image. Evaluation on synthetic data achieves an RMSE of 0.144 m and a depth accuracy (δ_1) of 87.3%. Additionally, the network demonstrates robust depth estimation in real underwater conditions, processing 256×256 images at 20 Hz. X. Yang *et al.* (2022) [195] propose a self-supervised depth estimation neural network designed to tackle underwater challenges. By incorporating image enhancement constraints, the network leverages distance variations in attenuated images, while a parallel optical flow network addresses issues of uneven lighting, noise, and moving objects. Image gradient constraints further refine texture details in the depth map, enhancing accuracy. Tests on real underwater data from the SQUID dataset demonstrate robust performance, with an average RMSE of 0.262 m, showing resilience even in blurry underwater conditions. A self-supervised architecture is also developed by J. Wang *et al.* (2023) [189] to improve depth estimation by synthesizing viewpoints through an iterative network that ensures precise relative pose estimation. This approach is further enhanced by an adaptive cost volume module, which assesses geometric compatibility across multiple generated viewpoints, allowing the network to effectively map pixel values to depth information. Evaluated on the Sea-thru monocular image dataset [196], the network achieves an RMSE of 0.415 m, a 70% improvement over the closest state-of-the-art model on this dataset. B. Yu *et al.* (2023) [197] present the UDepth network, which integrates a MobilenetV2 backbone [198], a transformer-based global attention module, and a convolutional regressor, optimized through a domain projection loss. UDepth’s tri-dimensional representations are evaluated on the USOD10K dataset [199] where it reports RMSE errors as low as 0.143 m.

rotated distillation strategy that improves rotational robustness. WaterMono achieves state-of-the-art performance in the FLSea-VI dataset [205] with an RMSE of 0.945 m and δ_1 depth accuracy of 89.2%. The architecture's generalization capabilities are explored with the FLSea-Stereo dataset where the performance of the model decreases to 1.616 m RMSE and 66.7% δ_1 .

Research considering deep learning for underwater scenarios has also been expanded to stereo techniques. K. Ichimaru *et al.* (2018) [206] mitigate the effects of the underwater floating particles through a CNN architecture that estimates depth from a stereo pair of images, while segmenting the disturbances caused by underwater phenomena to recover the original texture of the environment. The UWStereoNet introduced by K. Skinner *et al.* (2019) [207] is an unsupervised deep neural network that processes raw underwater stereo images to produce dense depth maps and color-corrected outputs. By incorporating an underwater image formation model, image processing techniques, and stereo vision's geometric constraints, the modular network achieves improved performance over a classic SGBM algorithm. X. Ye *et al.* (2023) [208] develop a novel underwater depth estimation model that adapts progressively to underwater conditions, leveraging synthetic underwater-style images from terrestrial datasets, and self-ensembling for semantic adaptation. Additionally, a disparity-range adaptation strategy addresses disparity mismatches, enhancing depth accuracy for close objects with large disparity variations. Q. Lv *et al.* (2024) [209] introduce UWStereo, a comprehensive synthetic underwater dataset containing image pairs with dense disparity annotations. The authors also incorporate masked image learning principles to develop a novel cross-view enhancement module. This module reconstructs cross-domain masks prior to stereo matching, improving the model's robustness and adaptability to various underwater scenes. The work of Q. Li *et al.* (2024) [210] proposes an unsupervised stereo matching approach that incorporates semantic information and attention mechanisms to guide the process. This method improves accuracy and overall performance, particularly in non-textured regions, demonstrating its effectiveness in challenging underwater environments where traditional stereo matching often struggles.

2.3.2 Perception-based Information Fusion

Measurements from sensors are inherently subject to various sources of uncertainty such as random errors or systematic biases that affect measurement accuracy in dynamic environments. Sensor fusion techniques help address these limitations by combining complementary data from multiple sources and/or modalities, creating single and more robust tri-dimensional representation of the scenario [39].

Heterogeneous Tri-dimensional Information Fusion

In aerial contexts, a large portion of the research concerning heterogeneous tri-dimensional data fusion focuses on combining stereo vision with LiDAR data since the information gathered from each of the sensors can be complementary. The first is textured and dense; while the second is very accurate and can be sparse or dense, depending on the resolution. W. Maddern and P. Newman (2016) [211] propose a real-time probabilistic approach capable of fusing depth maps

obtained via these sensors by propagating uncertainty estimates through a disparity prior and refinement stage. Tested on the KITTI dataset [212], this approach demonstrates high accuracy when data from both sensors is available, though performance noticeably declines when LiDAR data is missing. M. Ali *et al.* (2019) [213] propose a computationally efficient algorithm using multi-stage interpolation to densify non-uniform LiDAR samples, which are then dynamically fused with stereo-based information by assigning adaptive weights. This method achieves a 4.94% disparity error on the KITTI dataset, outperforming the classic SGBM algorithm, which shows a 13.56% error. A probabilistic framework that uses an Expectation-Maximization (EM)-based map matching methodology to fuse LiDAR information with Kinect 3D data is developed by Y. Yue *et al.* (2020) [214]. This approach has been tested in multiple real environments, demonstrating its versatility and accuracy in tri-dimensional mapping tasks.

Conventional and probabilistic methods often serve as a first processing step to learning-based algorithms, as these automatically extract features (both modality-wise and shared) directly from data [215]. A supervised learning approach to fuse this heterogeneous data is discussed by K. Park *et al.* (2018) [216]. The depth information of each sensor is processed separately at first, being fused at a later stage. This combined representation is then refined with features captured from a single RGB image. The authors report accurate results across multiple datasets, proving the generalization capabilities of the proposed network. Similarly, X. Cheng *et al.* (2019) [217] present an unsupervised LiDAR-stereo fusion network that addresses noisy data and sensor misalignments through a feedback loop mechanism. This approach further enhances learning by incorporating a piecewise planar model, allowing the network to more effectively represent the structured elements of the scene. When tested on the KITTI dataset, this model significantly improves the density of reconstructed 3D data, increasing LiDAR's original 7% coverage to a full 100% while also halving the error rate. The work of D. Martins *et al.* (2018) [218] employs a self-supervised learning strategy to fuse a stereo input point cloud with a dense depth prediction obtained through a CNN. The merging operation is a weighted interpolation of both depth values, enforcing preservation of high confidence stereo estimates while leveraging the monocular depth prediction in occluded regions. The resulting point cloud proves more accurate in experiments conducted with the KITTI dataset, as well as drone footage. Z. Bauer *et al.* (2019) [219] follows a similar principle, combining 3D camera depth readings with a deep learning-based monocular depth estimation. Point clouds are aligned, and a planarity measure is estimated for every point in the CNN-based prediction. The fusion operation occurs by comparing this measure to a given threshold: a higher ratio leads to maintaining the estimated point; otherwise the depth camera value is kept. This approach proves capable of outputting an improved point cloud for a set of very simplistic scenarios. The volumetric propagation network introduced by J. Choe *et al.* (2021) [220] leverages sparse LiDAR-derived geometric features to guide stereo matching correspondences within a unified 3D space. By embedding both stereo and LiDAR information into the same volumetric space, the network can propagate semantic and geometric information more effectively, leading to a reduction in depth estimation uncertainty. When compared to an early fusion method, the proposed intermediate fusion process shows improved metric results with over 2% gains in terms of RMSE.

H. Meng *et al.* (2023) [221] present FastFusion, a three-stage architecture designed to enhance 3D perception by efficiently integrating stereo and LiDAR data. In the initial feature extraction stage, a binary neural network captures high-level features from stereo images and the LiDAR map, using weighted distance metrics to improve stereo-matching reliability. The second stage implements an innovative cross-based LiDAR trust aggregation, which further boosts depth accuracy by combining LiDAR and image information during cost aggregation. Finally, a lightweight disparity refinement network takes the coarse disparity, reference image, and LiDAR data to deliver a refined and consistent output. Tested on several challenging datasets, FastFusion achieves a three-pixel stereo error of 2.81% with inference times averaging 16.8 ms. A holistic stereo-LiDAR fusion network that considers both intra- and inter-modality uncertainties to estimate an accurate depth map is introduced by J. Fan *et al.* (2024) [222]. This network is structured as a dual-branch architecture, where each branch processes stereo and LiDAR inputs separately with dedicated uncertainty modules to capture the unique uncertainty characteristics of each modality. A feature aggregation module then leverages these evidential uncertainties to fuse the outputs. Evaluation of the KITTI dataset shows a 6.52% RMSE improvement over the closest state-of-the-art network.

Combining heterogeneous information obtained from distinct imaging techniques is a rather unexplored area of the state-of-the-art. J. Shi *et al.* (2024) [223] propose a weighted fusion of stereo-based techniques with single laser line active imaging. The sparse point cloud estimated from laser triangulation is combined with the surface normals derived from the dense stereo data. This fusion minimizes an error function that optimizes the surface, incorporating both depth error and surface normal vector error. The fusion weights are adaptively calculated based on the angles between adjacent normal vectors, ensuring an optimized and detailed 3D surface reconstruction. Experimental results illustrated in figure 2.16 demonstrate the precision of the laser-based measurements, achieving an absolute error of 0.0349 mm. In contrast, the photogrammetric stereo data shows significantly higher absolute error at 0.9620 mm. Fusion results (depicted in figure 2.16c) achieve enhanced accuracy with an absolute error of 0.0293 mm, representing a 16% improvement over the laser-based technique and a 97% improvement compared to the passive stereo imaging method.

Literature regarding the fusion of heterogeneous tri-dimensional data in sub-sea environments is very scarce and generally guided towards sonar operations. J. Kim *et al.* (2019) [224] present a fusion technique that combines visual SLAM-derived maps with a sonar-based volumetric space-

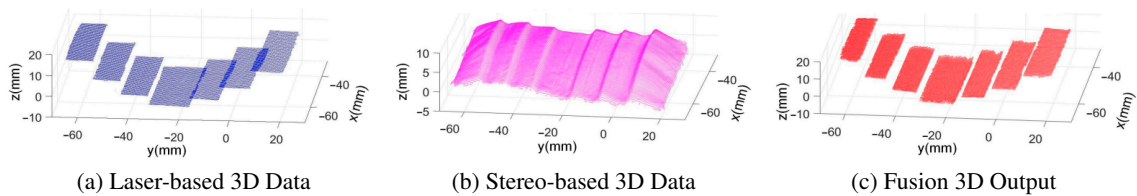


Figure 2.16: Sparse laser triangulation data is combined with surface normals from stereo methods, minimizing depth and normal vector errors with adaptively calculated weights. Extracted from [223].

carving model. The method integrates the elevation data from the optical model with depth measurements from sonar by examining boundary points and calculating their distances to a defined set of reference points. This approach improves the accuracy of 3D volumetric mapping by 50%, reducing mapping errors compared to traditional space carving alone. However, it requires significantly more computational power. A method for integrating 3D information from two sonars with orthogonal uncertainty axes is proposed by J. McConnell *et al.* (2020) [225], using a feature association technique. Features from each sonar output are clustered based on attributes such as range, intensity, and mean intensity kernel. To match features within each cluster, the method applies either a brute-force approach or a fast-convergence approach using random sampling to find the minimum-loss. In controlled water tank experiments, this system successfully inspects a blowout preventer structure, achieving a 3D reconstruction with an absolute error of 0.057 m. S. Negahdaripour (2020) [226] explores a stereo sonar configuration that utilizes epipolar geometry to reduce correspondence complexity. This process is refined through gradient descent optimization and validated in controlled scenarios, achieving errors between 0.02 and 0.10 m. H. Joe *et al.* (2021) [228] present a line-feature-based approach for enhancing 3D sonar maps by combining forward-looking sonar data with vertical scanning data from a profiling sonar. The two individual sonar submaps are matched with basis on critical points and corrected using the compensating information from each of the sensors. In pool trials, this method reduces slope and volumetric errors by 45.3% and 13.2%, respectively. Field tests at sea further validate the approach, showing that it reconstructs an aluminum cylinder with height and width errors of 12% and 5%. The features from overlapping areas between orthogonal sonars are also fused in the research of N. Sadjoli *et al.* (2023) [227] - see figure 2.17. To improve accuracy, the Smallest Of Cell-Averaging Constant False Alarm Rate (SOCA-CFAR) algorithm [229] is first applied to reduce acoustic noise in extracted features. The features are then matched across both sonar views by minimizing the loss between vectorized feature pairs. Finally, fused range, bearing, and depth data from each pair generate a uniform 3D representation. Controlled experimental results show that this approach yields

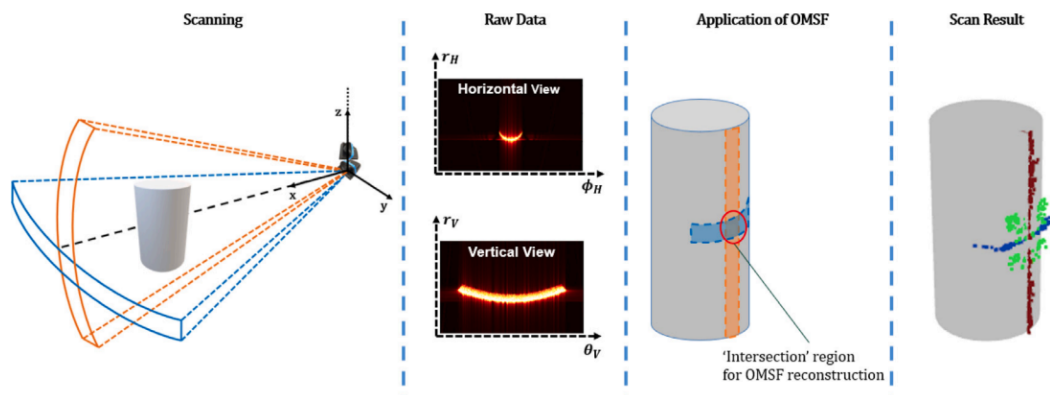


Figure 2.17: A configuration of two orthogonal multibeam sonars provide horizontal and vertical perspectives. Features from the overlapping regions can be matched to obtain an improved 3D representation of the target object. Extracted from [227].

higher mapping accuracy on solid surfaces, with slightly reduced precision for hollow objects.

A camera-sonar setup for underwater mapping and depth ratio estimation is proposed by A. Cardaillac and M. Ludvigsen (2023) [230]. Tri-dimensional acoustic data is matched to the closest image-based feature extracted from a SLAM algorithm [231]. The methodology was tested in two distinct pool environments, demonstrating distance and angle errors as low as 0.02 m and 0.15° , respectively. J. Zhang *et al.* (2024) [232] fuse sonar data with stereo vision by incorporating a sonar-based cost term directly into the SGBM algorithm. This integration refines the cost aggregation process during disparity estimation, leading to improved accuracy. Pool experiments demonstrate measurement errors down to 1.7%.

Multimodal Information Fusion

Sub-sea environments present significant challenges to acquiring high-fidelity 3D data. Fusing multimodal data such as textured 2D images, sparse depth and pixel motion cues; can mitigate these issues, enhancing the accuracy and overall perception of the underwater scene. The use of multimodal deep learning has been widely explored in aerial applications to estimate dense depth maps. An auto-encoder approach is proposed by C. Cadena *et al.* (2016) [233] to integrate RGB images, sparse depth cues, and semantic labels into a unified dense 3D representation. Each data modality undergoes an initial auto-encoding step for denoising, enabling the extraction of robust shared features. Depth estimation is subsequently performed from this shared feature space, leveraging the complementary information of all modalities. Validated on the KITTI dataset, this approach achieves real-time performance with inference times below 40 ms. M. Mancini *et al.* (2016) [234] leverage an encoder-decoder architecture for dense depth estimation using RGB images and optical flow information as inputs. The fusion of these modalities significantly improves performance, achieving over 9% gain on the KITTI dataset compared to a monocular depth estimation network. A residual network is introduced by Y. Liao *et al.* (2017) [235] fusing tri-dimensional information from a 2D laser with textured data. Sparse depth data is augmented by interpolating missing values along the gravity direction before being combined with RGB inputs at an early network stage. The final depth inference is refined by performing a pixel-wise sum with the augmented laser-based data, enhancing accuracy and producing a dense depth map. B. Ummenhofer *et al.* (2017) [236] shape the learning task to mimic a typical structure-from-motion problem. The DeMoN network is composed of two consecutive encoder-decoder architectures. Optical flow maps are first estimated and then fused with an RGB image pair to infer dense depth maps and surface normals. The Sparse-to-Dense network, proposed by F. Ma and S. Keraman (2018) [237], integrates RGB images with low-resolution depth data or sparse cues derived from SLAM. Experimental results on indoor datasets demonstrate that augmenting the input with 100 randomly sampled depth points increases the network's metric accuracy by over 50%. Dense depth predictions are achieved in the work of A. Hirata *et al.* (2019) [238] by interpolating missing values on LiDAR information. A propagation scheme that combines features from RGB images, semantic label and stereo-based to estimate depth in regions with missing data. The final output is further refined using a global optimization step. K. Wang and S. Shen (2020) [239] propose

a network that jointly estimates optical flow and camera motion from image pairs. A novel triangulation layer leverages this output, encoding motion information from multiple observations to achieve robust and accurate depth estimation. The research presented by P. N. Leite *et al.* (2020) [43] explores the effects of fusing RGB images with sparse depth cues from LiDAR, stereo matching, and active laser-based triangulation. Experimental results highlight the advantages of combining multimodal information over monocular depth estimation models. The study also reveals that models trained with sparse laser data exhibit limited spatial coverage, constraining their overall accuracy. D. Gehrig *et al.* (2021) [240] present a Recurrent Asynchronous Multimodal Network designed to integrate asynchronous, irregular event camera data with RGB images. Experiments on synthetic datasets demonstrate a mean absolute depth accuracy improvement of up to 30% compared to state-of-the-art networks.

A late-fusion approach is proposed by M. Khan *et al.* (2022) [241], employing a transformer network. Texture-based tokens are initially extracted from the input RGB image and subsequently fused with sparse LiDAR data, as well as the original RGB image, using a multimodal fusion block. Ablation studies reveal that this fusion methodology outperforms early fusion of RGB and sparse depth cues by approximately 19%. T. Shimada *et al.* (2023) [242] propose an early-fusion approach that combines an RGB image with optical flow information through an attention-based CNN. Additionally, a perceptual discriminator is employed to enhance the network's accuracy without increasing inference times. Experiments on the KITTI dataset demonstrate the metric benefits of incorporating motion-based data into the depth estimation task. The OccFusion network, proposed by J. Zhang *et al.* (2024) [243], introduces a multimodal framework combining LiDAR and multi-view image data. A 3D encoder processes LiDAR data to extract voxelized features, which are fused with point coordinates and image features from a 2D encoder. These modalities are integrated through a point-to-point fusion mechanism to generate enhanced voxel features. An active decoder further refines predictions in challenging regions, improving accuracy and robustness. B. Xiao *et al.* (2024) [244] develop an encoder-decoder network to fuse 2D RGB images with event camera data. The dual-branch multimodal encoder leverages self-attention mechanisms to correlate temporal and spatial information from texture patches and events. A joint decoder aggregates these feature maps using triple-scale feature fusion modules, enabling detailed and accurate depth estimation. A multi-stage framework is proposed by H. Li *et al.* (2024) [245] to fuse noisy Radio Detection And Ranging (RADAR) point clouds with RGB data. A first step generates an initial dense prediction from a monocular depth estimation task, followed by global-scale alignment using sparse RADAR points for refinement. A parallel branch learns associations between RADAR points and image patches to perform quasi-dense scale estimation. Finally, a scale map learner integrates the output of each branch for local scale refinement. Evaluation on the nuScenes dataset [246] demonstrates a 25.6% improvement in mean absolute error compared to the closest state-of-the-art network.

Multimodal information fusion in underwater environments typically focuses on integrating complementary visual and sonar data to enhance perception and mapping capabilities. N. Hurtós *et al.* (2010) [247] demonstrate a method that combines visual and sonar data by re-projecting

image textures onto tri-dimensional bathymetry models derived from multibeam sonar, merging camera-based visual detail with sonar range measurements. However, the operating range of the system remains restricted by the visibility limitations of the camera. C. Kunz and H. Singh (2013) [248] apply a similar approach by using image-based features to provide texture for multibeam sonar reconstructions. A pose graph is optimized to find the vessel's trajectory, camera pose, and resulting bathymetry map. Visual features are treated as landmarks, while partial sonar reconstructions are used for relative pose constraining. A large-scale mapping operation conducted by E. Iscar and M. Johnson-Roberson (2016) [249] fuses color-corrected visual data with multibeam sonar depth measurements to produce a textured, georeferenced bathymetry map with 15 cm resolution. A multi-layer architecture is proposed by D. Rao *et al.* (2016) [250] to correlate features from RGB images and *a priori* collected bathymetry data. This framework is further extended with gated multimodal learning techniques, enabling unsupervised clustering of input data and explicit learning of inter-modality features. M. Roznere and A. Li (2020) [251] fuse single-beam echosounder sonar data with SLAM-based monocular depth estimation. Measurements inside the sonar's detection cone are matched to visual features and integrated into the tracking phase of the SLAM algorithm to improve map initialization. This multimodal data fusion approach improves depth map accuracy and provides more reliable scale estimation.

An extension of this work is presented by H. Liu *et al.* (2023) [252], introducing a novel self-supervised network that combines RGB images with a masked representation of a single cone measurement from the same echosounder sonar (refer to figure 2.18). The acoustic range data is projected onto the image plane and intersected with foreground objects segmented from the textured data. Each input modality is processed independently and fused through a late-fusion depth decoder. The learning task is guided using real sonar measurements. Evaluation on a syn-

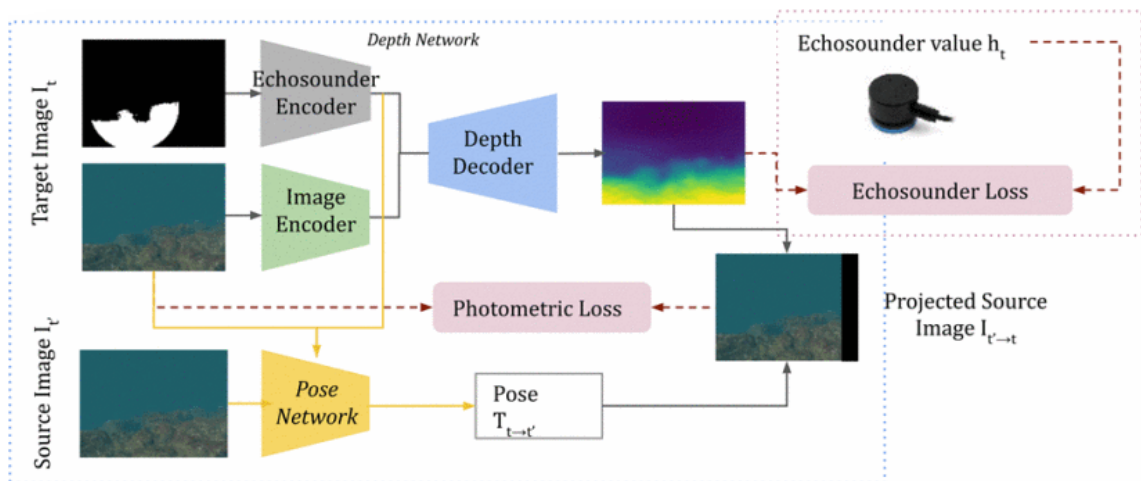


Figure 2.18: Self-supervised network that combines RGB images with masked echosounder cone measurements. Acoustic range data is projected onto the image plane and fused with texture-based features through a late-fusion depth decoder guided by real sonar measurements. Extracted from [252].

thetic dataset generated in a Unity⁶ simulation environment shows an average RMSE of 2.356 m and a δ_l pixel accuracy of 58.2%, outperforming other self-supervised methods. J. Zhang *et al.* (2024) [253] introduce a sonar-visual-inertial SLAM framework based on factor graphs. A dual-subsystem design ensures stable operation during visual failures, while joint feature-depth estimation from visual and sonar data enhances SLAM initialization and positioning accuracy. Additionally, a novel method for estimating sonar image elevation angles improves the stability and precision of the sonar-inertial subsystem. The UW_Depth network, proposed by L. Ebner *et al.* (2024) [254], addresses depth estimation and scale ambiguity by combining monocular RGB data with sparse depth triangulated from visual features. These sparse measurements are derived from ground truth values where valid pixel correspondences exist. The network architecture incorporates a lightweight backbone to process the coupled RGB-sparse prior input, a vision transformer for fusing output feature maps with original depth cues to capture global context, and a regression module that refines the depth map using a binning procedure. Evaluation on the FLSea-VI underwater dataset shows an RMSE improvement of 0.0083 m compared to using sparse measurements alone. Furthermore, the UW_Depth network achieves real-time performance on an embedded system, operating at 24 Hz.

2.3.3 Eye-in-Hand Perception

Advances in underwater intervention missions have enabled ROVs equipped with robotic manipulators to perform simplistic operations [255]. These systems typically rely on perception sensors mounted on the vessel, which can lead to occlusions caused by the robotic arm and increase risk to the vessel as it must approach the structure for detailed observations. Eye-in-hand configurations, where perception systems are coupled directly to the robotic arm, enhance situational awareness by taking advantage of the extra degrees-of-freedom. This setup enables data collection from previously inaccessible areas, minimizes occlusions, and provides close-range point-of-view information without compromising the vessel's safety. This configuration has been explored for fixed-based robotic manipulators in aerial applications, using monocular cameras [256], stereo vision systems [257], Red, Green, Blue, Depth (RGB-D) sensors [258], and structured light [259, 260].

Manipulation from floating platforms guided by an eye-in-hand perspective is a relatively unexplored topic in underwater perception literature. One of the first underwater intervention operations is presented in the research of G. Marani *et al.* (2009) [36], where the SAUVIM AUV is equipped with a 7 Degrees-Of-Freedom (DOF) electromechanical MARIS 7080 robotic manipulator, capable of sub-millimeter end-effector positioning precision (see figure 2.19). The system combines RGB images and ultrasound tracker feedback for target localization during a search and recovery mission. Within the manipulator's workspace, an eye-in-hand monocular camera detects and tracks the object during the hooking operation. Experimental trials require human intervention only to initiate target recognition, achieving translational detection accuracy within a centimeter and slightly higher errors in distance estimation to the target. A hybrid configuration is used by

⁶<https://unity.com/products/unity-engine>

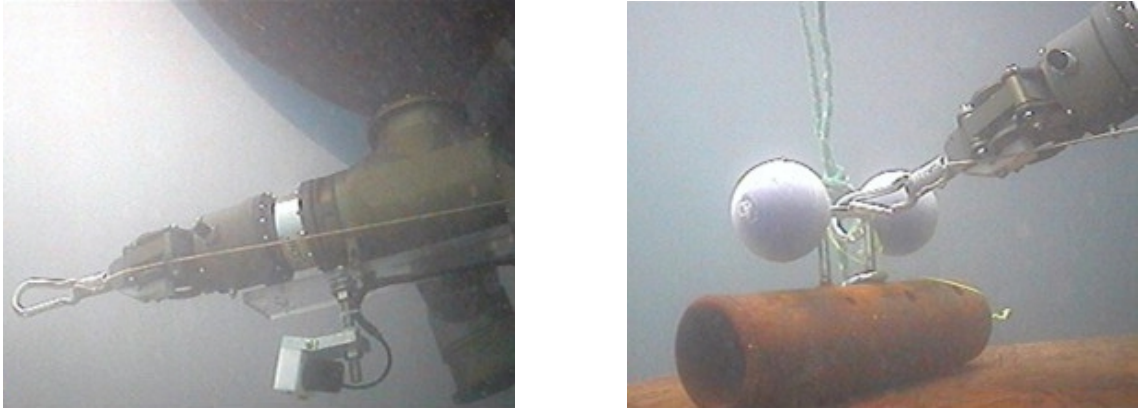


Figure 2.19: The SAUVIM AUV equipped with a 7-DOF MARIS 7080 manipulator uses an eye-in-hand monocular camera for object detection and tracking during a search and recovery mission. Extracted from [36].

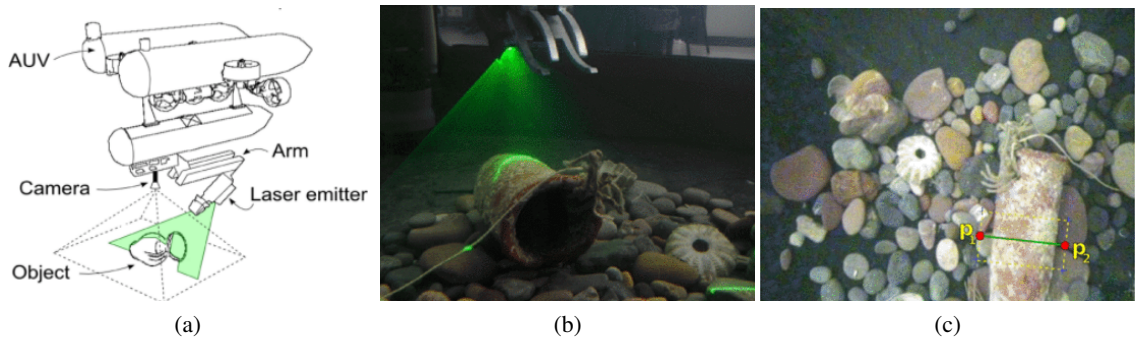


Figure 2.20: A hybrid eye-in-hand configuration. The laser diode is attached to the manipulator, while a monocular camera gathers image data from the AUV's perspective. Extracted from [261].

M. Prats *et al.* (2012) [261] for grasping operations. A laser emitter is mounted on a 4-DOF Light-weight ARM 5E manipulator, while a downward-facing camera is integrated into the AUV (figure 2.20a). The scene is reconstructed by moving the manipulator's elbow joint at a constant speed, generating sparse point clouds through triangulation of multiple laser stripe scans (figure 2.20b). These point clouds are stitched together using the motion of the floating platform. A human operator determines the optimal grasping position (figure 2.20c), after which the grasping plan is executed autonomously. M. Carreras *et al.* (2015) [65] present a custom end-effector with an embedded camera and force-torque sensors for underwater valve-turning interventions. The AUV uses a Bumblebee 2 stereo system to detect and position itself near the target. The grasping operation is guided by an eye-in-hand vision system, which measures orientation errors between the valve and the end-effector. The system achieves an 87.5% success rate, with failures attributed to manipulator calibration errors or inaccurate valve pose estimation. A single-stripe structured light vision system is attached to a manipulator in the work of A. Peñalver *et al.* (2015) [37]. Tri-dimensional reconstructions of the scenario are obtained as the perception system moves along with the robotic arm. Experiments in the UWSim [262] simulator demonstrate that the eye-

in-hand configuration provides more complete and accurate reconstructions compared to a camera mounted as part of the AUV payload. R. Conti *et al.* (2017) [263] simulate a 7-DOF robotic arm equipped with a three-fingered gripper and an eye-in-hand camera mounted on the end-effector. The AUV aligns with the target object, and the grasping plan is guided by the estimated center of mass and orientation of the target object relative to the manipulator, as determined by the visual camera.

Eye-in-hand vision-based control strategies are employed by S. Sivčev *et al.* (2018) [264] to guide intervention missions with the work-class Holland I ROV. A Point Grey Blackfly⁷ camera is integrated with the Schilling Titan 2 subsea manipulator, achieving positioning errors as low as 15 cm. Controlled experiments and real-world trials validate this approach across various tasks, including precise arm control near a fiducial marker and grabbing and rotating a T-bar target handle. J. Li *et al.* (2019) [265] also propose a control system that uses a monocular camera mounted on the manipulator wrist. The algorithm, validated through numerical simulations, addresses uncertainties in camera parameters, vehicle kinematics, and feature position information to enhance control accuracy and reliability. A hybrid approach for underwater inspection is introduced by M. Santos *et al.* (2021) [266], combining complementary visual data from an ROV-mounted camera and an eye-in-hand camera. A switching mechanism adjusts viewpoints during operation, activating the ROV camera for broader scene awareness when the target leaves the eye-in-hand camera's field of view. Real-world tests demonstrate the effectiveness of this system, with the eye-in-hand configuration enhancing close-range inspections by reducing positioning errors, while the ROV camera provides valuable contextual information. J. Gao *et al.* (2021) [267] use a monocular camera to extract texture-based features, which are input into a model predictive visual servoing system. This approach enables a decoupled control of the ROV-manipulator system.

A cooperative intervention mission is discussed in the research of R. Pi *et al.* (2021) [268], using two Girona500 vessels equipped with 4 DOF ECA 5E manipulators. The shared transportation of a target pipe is achieved through the use of custom cameras embedded in each of the end-effectors (see figure 2.21). These cameras provide visual feedback for detecting and segmenting a target pipe, with grasping points estimated from the eye-in-hand perspective. G. Billings *et al.* (2022) [269] present a hybrid SLAM approach that combines a stereo camera mounted on an ROV with an eye-in-hand monocular fisheye camera. The dynamic positioning of the wrist-mounted camera is leveraged to extend the initially constrained map, improving spatial coverage. Underwater experiments highlight the advantages of integrating multiple camera viewpoints. However, the algorithm does not account for the robotic manipulator's current state, which limits the robustness of the resulting 3D representation. A hybrid configuration for visual servoing control in underwater vehicle-manipulator systems is presented by J. Li *et al.* (2024) [270] (refer to figure 2.22). This approach employs position-based visual servoing using a stereo camera mounted on the ROV, providing 3D information and a broad FOV. Additionally, an uncalibrated visual servoing method uses visual feedback from an eye-in-hand camera to guide manipulator control during

⁷<https://www.teledynevisionsolutions.com/products/blackfly-s-usb3/>

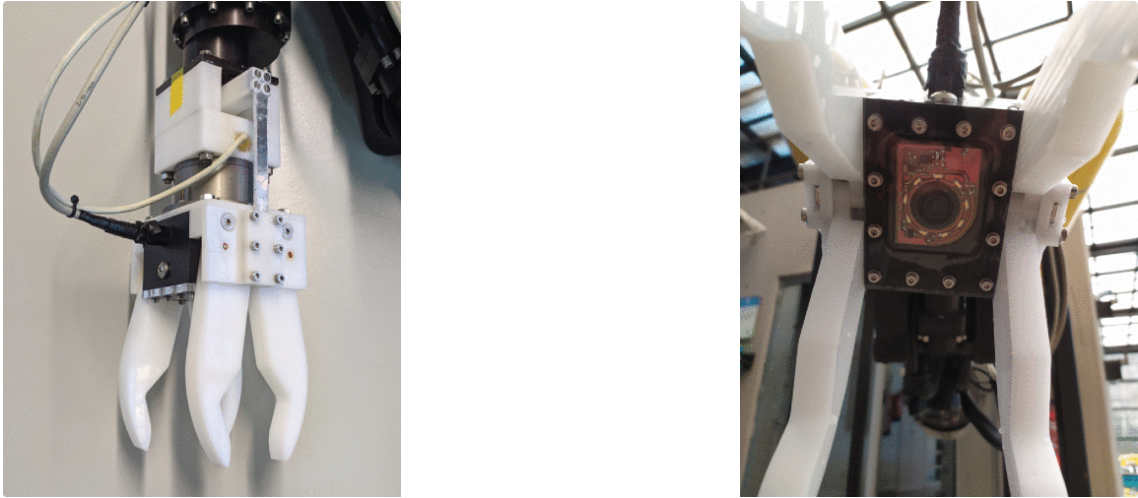


Figure 2.21: Custom end-effector gripper and camera to enable visual pipe detection, segmentation, and grasping point estimation from an eye-in-hand perspective. Extracted from [268].

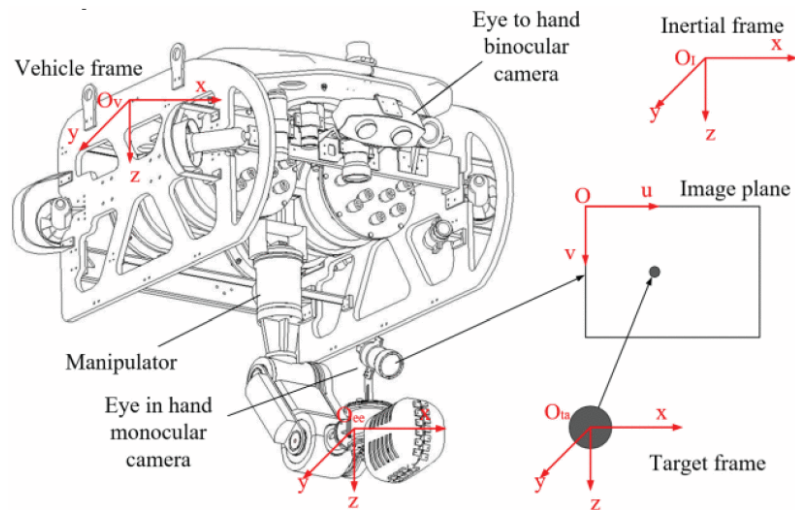


Figure 2.22: Hybrid visual servoing configuration for underwater vehicle-manipulator systems, combining an ROV-mounted stereo camera and an eye-in-hand camera. Extracted from [270].

close-proximity operations. Validation through simulations and field trials demonstrates the system's effectiveness in moving toward a static target in simplistic scenarios.

2.4 Critical Analysis

Underwater perception systems face significant challenges from sub-sea phenomena that lead to severe data degradation. Mitigation strategies include post-processing techniques that reduce haze, blurriness, and color distortion; significantly improving feature-matching algorithms by sharpening textured details in captured images. Additionally, the impact of floating particles can be reduced during image formation by gating the acquisition process to avoid a portion of the backscat-

tered photons. This method requires precise synchronization between the camera shutter and the projection of artificial lights onto the scenario. Compared to traditional approaches that rely on constant illumination — resulting in degraded image quality due to persistent backscattering, overexposure, and vignetting effects; range-gated imaging significantly enhances the perceptual capabilities of underwater vision systems. The design of waterproof housings for underwater imaging systems significantly affects the quality of the retrieved visual information. Flat port surfaces have limited FOV, with radial distortions increasing proportionally to the distance from the camera’s center of projection. In contrast, hemispherical dome ports effectively minimize light refraction when the optical axis is aligned with the dome’s center. The calibration of underwater optical systems remains an open-research problem with literature advancing towards complex mathematical ray tracing models powered by learning-based models. In comparison, radial distortion correction methods account for the combined effects of lens distortion and refraction at the housing interface, eliminating the need for explicit modeling of complex optical paths through multiple media. Ad-hoc underwater calibrations further enhance accuracy by capturing the actual refractive conditions that account for variations in water properties.

The state-of-the-art in underwater perception is extensive; however, the information obtained by these solutions often lacks the accuracy required to perform thorough tasks that require millimetric precision. Table 2.1 provides an overview of the most relevant imaging techniques and sensors. Sonar systems excel in long-range mapping due to the low attenuation of sound underwater. Despite this advantage, sonar technology is constrained by its limited resolution and struggles with fine, close-range details due to overlapping echoes and navigation drifts. Optical systems are versatile and capable of capturing data with distinct characteristics. Passive stereo imaging methods generate dense and textured representations but are highly susceptible to degradation due to light scattering and low-texture regions. Alternatively, 3D information can be derived from a single moving camera using photogrammetry principles. While SfM methods are effective for reconstructing 3D scenes, these rely on precise camera tracking, perform poorly in dynamic environments, and lack the ability to recover the absolute scale of the scene without external information. Active imaging techniques offer high accuracy in describing volumetric dimensions, but provide sparse information that is insufficient for a complete perception of the scene without sweeping motions. These systems are highly sensitive to lighting variations, which can disrupt segmentation and triangulation processes. Recent advancements incorporate sequentially adjustable mirrors to improve spatial coverage, at the expense of larger housings and lower acquisition ratios. Pattern projectors enhance scene texture, however, their high power consumption restricts deployment in real-time or energy-constrained missions. Turbid conditions degrade the effectiveness of these techniques due to absorption. Learning-based solutions demonstrate potential for highly accurate and dense reconstructions, as generative models start to be employed in underwater environments for style adaptation. The widespread adoption of these approaches is constrained by the availability of large-scale datasets, the absence of a standardized platform for fair performance evaluations, inaccuracies in ground-truth distance annotations, and a lack of diversity in scenarios. Hybrid imaging systems acquire both dense, textured 3D information from passive methods, and

Table 2.1: Advantages and disadvantages of distinct sensors and techniques used for underwater tri-dimensional perception.

Technique/Sensor	Strengths	Weaknesses	Data Characteristics
Sonars	Long-range measurements. Resistant to water turbidity and light conditions.	High minimum operation distance. Sensible to media density variations. Multipath propagation echoes.	Low resolution Non-textured Sparse
Photogrammetric Stereo	Close-range measurements. Color mapping for each point. High spatial coverage and resolution.	Depends on accurate feature matching. Sensible to sub-sea phenomena. Depth precision decays with acquisition distance.	Textured Blocky Dense
Structure-from-Motion	Cost-effective. Close-range measurements. Color mapping for each point. High spatial coverage and resolution.	Depends on accurate feature matching. Requires accurate camera motion tracking. Susceptible to dynamic environments. Sensible to sub-sea phenomena. Scale ambiguity.	Textured Blocky Dense
Learning-based	High potential for accurate measurements. Learns directly from raw data. Able to estimate depth from a single image.	Computationally demanding. Requires large amounts of training data. Generalization difficulties.	Textured Very Dense
Structured Light (Laser Stripe Triangulation)	Close/Medium-range measurements. Easily extended to multiple laser configurations. Millimeter-precise depth information.	Depends on accurate laser segmentation. Absorbed quickly in turbid environments. Limited spatial coverage.	High resolution Non-textured Sparse
Structured Light (Pattern Projectors)	Creates artificial texture. Improves stereo correspondences. Variety of available patterns.	High power consumption. Absorbed quickly in turbid environments.	Dense

millimeter-precise depth measurements from laser triangulation. These are complex to assemble and require precise calibration and synchronization between imaging modes.

The inherent error characteristics of individual techniques or sensors can be mitigated by fusing multiple data sources. In aerial robotics applications, tri-dimensional data fusion typically involves one-to-one mapping of dense point clouds, using probabilistic methods or deep learning architectures. However, the lack of dense and precise 3D data makes these methodologies unfeasible in sub-sea environments. Furthermore, strategies designed to handle missing data often rely on refinement stages heavily dependent on RGB-based features, limiting their applicability underwater. Existing literature on fusing tri-dimensional data from distinct underwater imaging techniques is limited, often resorting to simplistic weighted averaging of point clouds, which fails to fully exploit the complementary strengths of each input. State-of-the-art research heightens the advantages of combining multimodal data. Texture-based features from RGB images are often combined with sparse depth data from LiDAR or RADAR sensors for depth completion. Additionally, incorporating pixel motion information into learning architectures demonstrates improved depth map accuracy, as motion cues encode valuable depth-related features. Heterogeneous 3D information fusion for underwater environments focuses on combining overlapping acoustic data from orthogonal sonar configurations. On the other hand, multimodal fusion algorithms in sub-sea contexts commonly integrate texture data with sonar measurements for bathymetry effects. Tri-dimensional reconstructions from RGB and sparse depth priors are also discussed in the literature, leading to improved results when compared to underwater monocular depth estimation approaches. Nevertheless, these sparse depth priors are often randomly sampled from available ground truth data, introducing bias into the learning process and limiting the generalization capabilities of the models.

Autonomous underwater interventions typically rely on vessel-mounted perception systems,

which are prone to occlusions and lead to manipulation accuracy. Eye-in-hand configurations address these limitations by coupling the perception system directly to the manipulator. This approach leverages the arm's additional DOF to access previously unreachable areas, minimize occlusions, and enable close-range inspections. State-of-the-art eye-in-hand applications are limited to monocular cameras for optical feedback and vision-based control. Structured light imaging, in contrast, often employs a hybrid configuration where a laser projector is mounted on the robotic arm while a vessel-mounted camera captures the images. This setup improves spatial coverage by using the arm's movement for sweeping motions, eliminating the need for complex galvanometer-based systems. Nevertheless, the images captured from above still face occlusions from the arm itself. Recent advancements in eye-in-hand technology focus on hybrid configurations with dual perception systems. A vessel-mounted sensor provides a broader field of view for navigation and coarse reconstructions, while the eye-in-hand system focuses on close-range inspections, offering complementary perspectives for enhanced situational awareness.

The research presented in this doctoral thesis builds upon the state-of-the-art by exploring hybrid imaging techniques. The MARESy sensor is designed to minimize distortions through hemispherical dome ports and thorough underwater calibration. Additionally, the effects of floating particles are reduced through a data-driven range-gated imaging scheme. Novel information fusion algorithms are introduced to tackle the fusion of heterogeneous 3D point clouds from each imaging technique, diverting from state-of-the-art approaches by propagating the sparse 3D information towards the stereo-based representation. Furthermore, a multi-stage approach draws from learning-based multimodal fusion principles to combine textured data, the apparent magnitude of pixel motion, and sparse depth measurements into a single dense representation of the submerged scene. In essence, MARESy is a self-reliant imaging sensor capable of gathering high-fidelity information by resorting to mitigation strategies and information fusion techniques that integrate complementary data. Eye-in-hand configurations using hybrid imaging can fill the research gap when it comes to gathering millimeter-precise tri-dimensional information for the means of thorough O&M inspections.

Chapter 3

MARESyE: Hybrid Imaging for Tri-dimensional Underwater Scene Understanding

3.1 Introduction

The submerged portion of O&M structures pose a distinct set of challenges, as the elevated level of salinity in seawater aggravates the intensity of corrosion. Cathodic protection systems such as sacrificial anodes are usually employed to degrade in place of the steel structure. As a result, regular inspection operations are needed to assess the anode regarding its state of corrosion. Inspection procedures aim to detect defects (e.g. cracks, buckling, corrosion), assess the thickness reduction of the anode and the condition of its coating, and finally, detect structural deformations [12].

Assessment of these structures can be conducted through direct contact with the anode, using direct current densities and half-cell or open-circuit potential measurements [271, 272, 273]. These tests require the detachment of the anode from the main structure, leading to extended downtimes [274]. Magnetic Particle Inspection (MPI) can detect small cracks underwater but require divers due to automation challenges and the need to remove protective coatings [12]. Non-invasive approaches such as Alternating Current Field Measurements (ACFM) use probe-induced electric currents to measure electromagnetic field disturbances near the surface, detecting anode defects [275]. Similarly, Eddy Current (EC) inspection provides surface crack detection and measurement [276]. Visual inspections range from quick overviews of large-scale defects and assembly verification to thorough examinations of cracks, corrosion, and biofouling. These vision-based assessments are cost-effective in comparison to traditional methods [12], but currently rely on divers with handheld cameras — a dangerous and unreliable process. While ROVs and AUVs offer safer alternatives, these lack the situational awareness for close-proximity operations [255]. This limitation primarily stems from inadequate tri-dimensional perception of the surrounding environment. Traditional solutions using off-the-shelf vision systems in watertight enclosures [128]

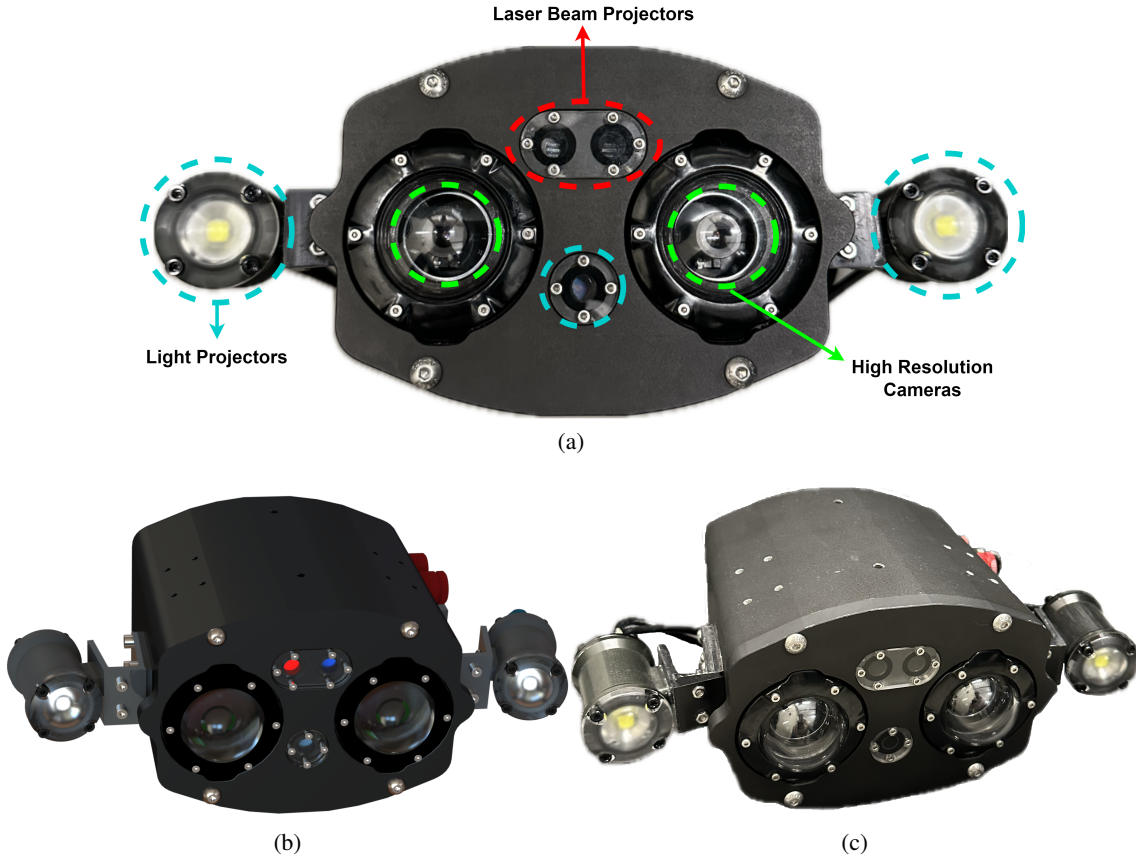


Figure 3.1: The MARESyE hybrid imaging sensor. a) The system employs a pair of high resolution cameras for stereo vision, two laser beam projectors that allow for active imaging techniques, and multiple light source projectors. A triggering system synchronizes data acquisition via hardware, while an internal processing unit enables edge computing. b) Isometric perspective of the simulated render of the sensor. c) Isometric perspective of the final prototype of the sensor.

suffer severe data degradation from underwater phenomena like light absorption, scattering, and suspended particles [44].

The research presented in this chapter details the MARESyE imaging sensor, designed to retrieve high-fidelity information from close-range underwater environments. MARESyE can be used as ROV payload to remotely guide a visual inspection operation by providing millimeter-precise reconstructions. This allows for a metric assessment of the sacrificial anode’s wear, as well as the size of potential cracks and other defects. The MARESyE device is based on a patented hybrid imaging concept¹ that enables multimodal information (textured images, tri-dimensional information, motion-based data) to be retrieved from harsh sub-sea environments [87, 38]. Figure 3.1 illustrates both its simulated rendering and the finished prototype. The remainder of this chapter is organized as follows: section 3.2 presents MARESyE’s hybrid imaging principles and data-driven formulation, while sections 3.3 and 3.4 introduce the optical and mechanical design considerations that enable high-quality underwater imaging. Section 3.5 provides a comprehensive

¹This technology is patented under: US11503269B2 (granted); PCT/IB2019/052926, EP3775996, AU2019251438A1 (pending).

performance analysis through experiments in both controlled environments and real inspection operations at the ATLANTIS Coastal Test Center. Finally, section 3.6 summarizes the key findings and implications of this chapter.

3.2 Hybrid Imaging and Data-driven Formulation

A hybrid imaging sensor leverages both passive and active techniques. Each methodology is triggered separately and sequentially using a data-driven formulation. MARESy employs Photogrammetric Stereo (PS) principles to retrieve dense and textured 3D information. Additionally, laser beams are projected onto the scenario enabling precise triangulation of sparse information through Light Stripe Ranging (LSR). The stereo-based algorithm derives a tri-dimensional understanding of the environment from multiple camera viewpoints. Photogrammetry-based principles are used to calculate pixel-to-pixel displacements between the left and right images - disparity estimation. These pixel correspondences are estimated using the SGBM algorithm [140]. The established optical baseline enables the triangulation of depth information as the inverse of disparity. The tri-dimensional information obtained through this method is dense and textured, however, it

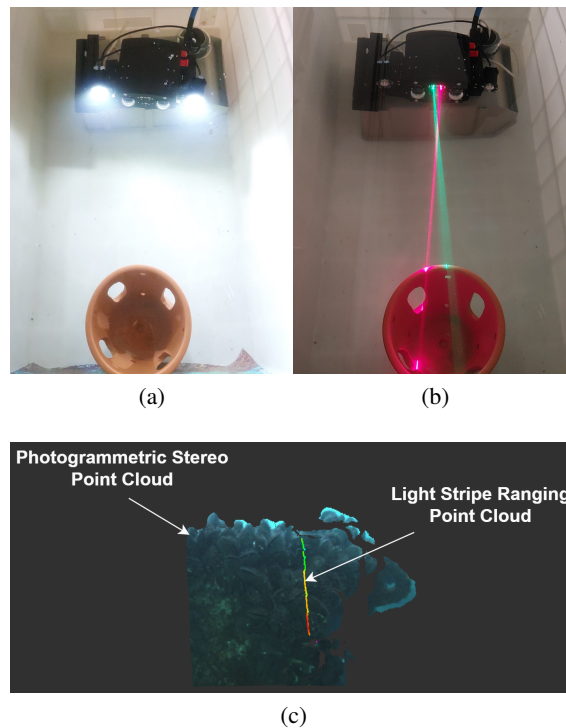


Figure 3.2: The photogrammetric stereo stage is depicted in figure a) - light projectors are turned on to illuminate the scene, while laser diodes are off. Figure b) portrays the light stripe ranging acquisition - laser diodes are projected onto the scene, while the high intensity LEDs are turned off to facilitate the segmentation procedure. The heterogeneous tri-dimensional information gathered by the MARESy sensor is illustrated in figure c).

heavily relies on environmental features and consistent illumination conditions. The use of external Light-Emitting Diode (LED) sources facilitates the operation of the PS module regardless of the ambient light conditions - refer to figure 3.2a. On the other hand, the LSR technique resorts to projecting high intensity laser beams that actively modify the observed scenario - figure 3.2b. Segmentation of each beam is conducted to enable the individual processing of each laser. The remaining pixels are triangulated into tri-dimensional information through a previously estimated camera-laser calibration plane. Each beam is correlated into 3D information relative to its respective camera. The information derived from LSR is highly accurate in depicting the spatial and volumetric dimensions of objects in the scene. However, it consists of a sparse set of points that strongly depend on the quality of the segmentation process. The point clouds resulting from MARESyE's PS and LSR imaging techniques are illustrated in figure 3.2c.

The aggregation of organic material and debris in the water deeply influences its turbidity and, consequently, the distance that light can propagate through. In addition to this fact, floating particles are also the cause for light to be deviated from its path. The portion of light that gets reflected back to the imaging sensor - backscattering - vastly hinders visibility and induces contrast limitations [83]. MARESyE addresses this issue through a range-gated system that controls the image formation process [87]. The start of the image acquisition process ($t_{open_shutter}$) is delayed by a defined t_{delay} relative to the activation of the external lights or lasers. The camera shutter is precisely timed to avoid capturing the backscattered light from the nearby marine snow. It remains open for a fixed exposure time ($t_{exposure}$) to accumulate light reflections containing relevant scene information. Figure 3.3 illustrates this time-gated image formation process. This system serves as a high-level control unit, sequentially gating each module individually to ensure that each im-

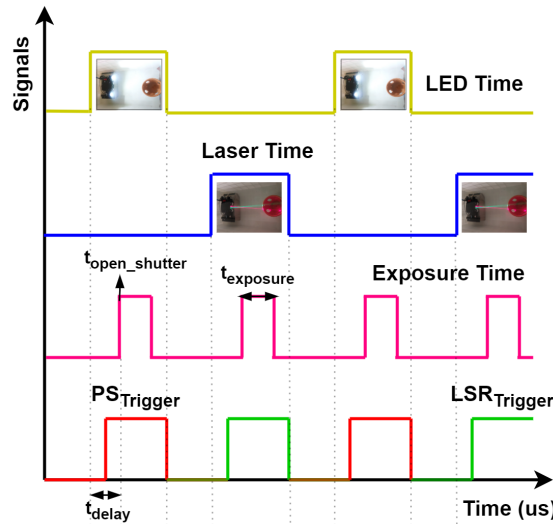


Figure 3.3: A range-gated synchronizer is employed by MARESyE to mitigate the impact of the floating particles. This strategy allows for the imaging sensor to receive and accumulate light reflections from the scenario, while avoiding backscattered photons reflected from debris that sit close to the observer.

age remains unaffected by concurrent modes. The capturing rate of the sensor is driven by this hardware triggering mechanic, guaranteeing synced 2D imaging at 30 Hz (15 Hz for each mode). MARESyE then processes the captured raw image data internally resorting to its embedded processing unit, estimating tri-dimensional information at 10 Hz and 5 Hz for the LSR and PS modes, respectively. In essence, this system allows MARESyE to function in harsh underwater environments (high levels of sediments in the water, no ambient light, etc); to reduce power consumption and improve the accuracy of the synchronization process; and to mitigate the impact of motion blur introduced when the sensor is integrated within underwater vehicles [87].

3.3 Optical Design for High-quality Underwater Imaging

The MARESyE sensor is tailored for the inspection and maintenance of submerged critical structural elements in offshore platforms. This application requires very thorough and precise measurements. MARESyE is designed to gather millimeter-accurate tri-dimensional information in close proximity of the target surface, with an optimal workspace around the 0.6 m distance. Processing the gathered information on the edge using MARESyE's internal processing unit can provide the real-time volumetric feedback that is crucial for guiding the inspection procedure, allowing for a continuous assessment of the submerged structure. The core of the optical system relies on two high-resolution 1440×1080 px cameras that allow for highly flexible parameterization. The Sony CMOS Pregius IMX273 image sensor features a pixel size of ($H_{pixel\ size} \times V_{pixel\ size} = 3.5 \times 3.5 \mu\text{m}$) and a dynamic range of 10 bits. Its fast response to external triggering (down to 1 μs) and the use of a global shutter mechanism enable the range-gated system, ensuring the capture of high-fidelity information directly during the image formation process. Each camera is equipped with a 6 mm focal length lens, focused to provide a depth of field that ensures objects within MARESyE's desired operational range of approximately 0.6 m remain in focus. The FOV for camera-lens optical configuration can be estimated through equations 3.1 and 3.2:

$$FOV_{horizontal} = 2 \cdot \arctan\left(\frac{H_{pixel\ size} \cdot NPixel_{width}}{2 \cdot f}\right) = 44.99^\circ, \quad (3.1)$$

$$FOV_{vertical} = 2 \cdot \arctan\left(\frac{V_{pixel\ size} \cdot NPixel_{height}}{2 \cdot f}\right) = 34.53^\circ, \quad (3.2)$$

where $H_{pixel\ size}$ and $V_{pixel\ size}$ correspond to the sensor's pixel size; the width and height of the captured image are represented as $NPixel_{width}$ and $NPixel_{height}$, respectively; and f represents the focal length of the optical lens.

A comprehensive understanding of the spatial resolution of MARESyE's PS point clouds can be obtained from equations 3.3 and 3.4 [277]:

$$SpatialResolution_x = \frac{2 \cdot Z \cdot \tan\left(\frac{FOV_{horizontal}}{2}\right)}{NPixel_{width}}, \quad (3.3)$$

Table 3.1: Evolution of MARESyE’s PS point cloud spatial resolution in function of Z.

Z (m)	Point Cloud X and Y Spatial Resolution (m/pixel)	Z (m)	Point Cloud X and Y Spatial Resolution (m/pixel)
0.376	0.00022	0.726	0.00042
0.426	0.00025	0.776	0.00045
0.476	0.00027	0.826	0.00048
0.526	0.00030	0.876	0.00050
0.576	0.00033	0.926	0.00053
0.626	0.00036	0.976	0.00056
0.676	0.00039	1.026	0.00059

$$SpatialResolution_Y = \frac{2 \cdot Z \cdot \tan\left(\frac{FOV_{vertical}}{2}\right)}{NPixels_{height}}, \quad (3.4)$$

where Z translates the metric depth of the scene.

The spatial resolution of the 3D point cloud depends on the distance from which the information is retrieved. Table 3.1 shows this resolution across a metric range of [0.376, 1.026] m, demonstrating that MARESyE provides sub-millimeter point cloud spatial resolution throughout the complete operational range, with precision as high as 0.00022 m.

The internal design of the sensor is optimized to achieve the best performance in close-range operations while keeping the housing as compact as possible. A baseline of 0.084 m between cameras is enforced for the stereo configuration. Additionally, external 1700 lm light sources, crucial for this methodology, are attached externally to the housing, deviating 10° from the camera plane. This arrangement ensures consistent illumination across MARESyE’s field of view, with the maximum light incidence estimated for the 0.65 m distance – see figure 3.4a. MARESyE’s active technique makes use of two laser diodes: a 200 mW, 650 nm (red) wavelength beam; and a 300 mW, 405 nm (blue-violet) wavelength. A compromise between maximum power and rising time is achieved to create a synergy that allows for the image formation to capture the brightness peak of the laser beam, while actively responding to the triggering signals. The resolution of the LSR technique relies on the interception between the laser beam projection and image planes. The optical setup is thought out to meticulously align each laser diode with the opposite camera at a specific operational distance. As depicted in figure 3.4b, a 5° deviation achieves plane alignment at the 0.60 m distance. This internal disposition serves as an effective compromise between optical performance and housing design considerations: a compact design that enables the optical prowess of the sensor.

The fidelity of the retrieved information can also be affected by deviations induced while light traverses media with different optical densities [278]. The optical housing infrastructure has direct impact in mitigating these refractions. A flat housing port causes light rays to intersect in distinct points, leading to multiple apparent viewpoints [97]. The hindrances caused by refraction can be severely reduced if a hemispherical interface is employed. Aligning the center of the optical system’s projection with that of the interface ensures that the refracted light converges at the

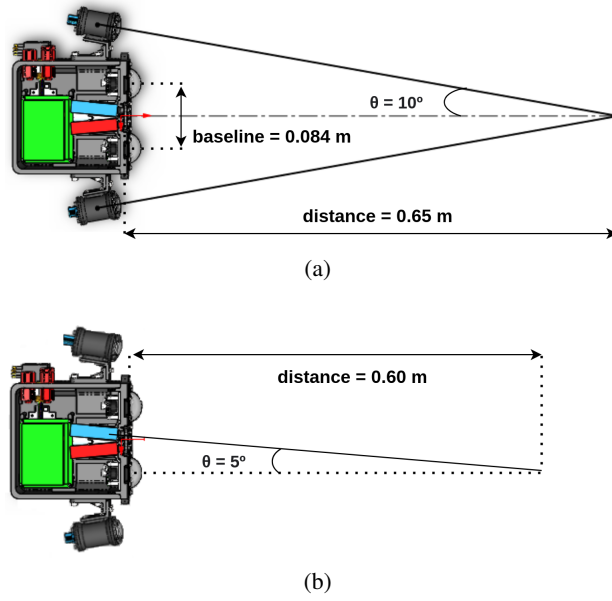


Figure 3.4: Internal design implications on the optical setup of the MARESyE sensor. Figure a) depicts the stereo-based configuration with a camera baseline of 0.084 m. The external lights ensure constant illumination of the imaging workspace with maximum incidence at the 0.65 m distance. Figure b) shows that a 5° LSR plane deviation creates a perfect camera-laser alignment in the 0.60 m distance range.

same viewpoint [99]. Moreover, dome-shaped structures offer greater stability and can withstand high depth-rated pressure environments. Nevertheless, a flawless assembly is mechanically difficult to achieve due to small alignment deviations or even existent non-uniformities in the glass. MARESyE makes use of hemispherical housing ports to accommodate the optical system, physically reducing light refractions. The calibration of optical systems in underwater environments remains an open research problem, with state-of-the-art approaches ranging from geometric modeling to the implicit absorption of underwater optical effects using the pinhole camera model - refer to section 2.2.2.

MARESyE's calibration process employs a radial distortion correction approach to effectively model lens and dome distortions, along with minimal mechanical-induced deviations [99]. The sensor's lens are specifically focused for the designed operating range and fixed within a clear water tank, while a 9×12 chessboard with 0.015 m sized squares is used for both intrinsic and extrinsic camera calibrations. The global Levenberg-Marquardt optimization algorithm minimizes the re-projection error of the chessboard points [279]. The camera-laser planes for LSR triangulation are calibrated in the same controlled environment, following the methodology of A. Pinto and A. Matos (2020) [87]. The calibration setup employs a chessboard with an elevated side-piece, where laser beams are projected across both surfaces. The spatial arrangement between each laser and camera frame is determined by estimating the plane that best fits the segmented laser projection points. A stereo re-projection error of 0.186 px after calibration further supports the choice of the pinhole camera model for mapping tangential and radial distortions induced by both the

underwater environment and the camera-lens-dome optical setup. Following this camera model, the depth of the scene Z can be obtained from equation 3.5:

$$Z = \frac{f \cdot B}{d}, \quad (3.5)$$

where f and B correspond to the focal length and stereo baseline, respectively; d translates the estimated disparity value.

The re-projection error in stereo vision manifests as an error in the disparity Δd , which leads to uncertainty in the estimated depth ΔZ . An approximation of this error can be quantified by taking the derivative of the depth equation with respect to disparity - equation 3.6:

$$\frac{\partial Z}{\partial d} = -\frac{f \cdot B}{d^2}. \quad (3.6)$$

The ΔZ uncertainty can thus be approximated through equation 3.7:

$$\Delta Z \approx \left| \frac{\partial Z}{\partial d} \right| \cdot \Delta d = \frac{f \cdot B}{d^2} \cdot \Delta d. \quad (3.7)$$

Substituting in the disparity d from equation 3.5, the uncertainty introduced by the re-projection error in the depth estimation of the PS point cloud can be quantified using equation 3.8 [108]:

$$\Delta Z = \frac{Z^2}{f \cdot B} \cdot \Delta d. \quad (3.8)$$

This uncertainty increases quadratically with distance from the sensor. Table 3.2 shows the evolution of ΔZ for the studied range. For the designed operating range of 0.6 m, MARESyE's stereo-based depth estimation has an associated uncertainty of 0.00033 m. Sub-millimeter ΔZ values can be expected for the [0.376, 0.826] m spectrum. For larger distances, the depth uncertainty increases to the millimeter order of magnitude.

Table 3.2: Evolution of the uncertainty in MARESyE's PS depth estimation in function of Z .

Z (m)	Uncertainty (m)	Z (m)	Uncertainty (m)
0.376	0.00019	0.726	0.00070
0.426	0.00024	0.776	0.00080
0.476	0.00030	0.826	0.00091
0.526	0.00037	0.876	0.00102
0.576	0.00044	0.926	0.00114
0.626	0.00052	0.976	0.00127
0.676	0.00061	1.026	0.00140

3.4 Mechanical Design and Validation

The mechanical design of underwater components must account for various factors dictated by the sub-sea environment. The corrosion of metallic components is an important concern due to the

structural degradation caused by the salinity of seawater. Ensuring a waterproof housing is essential to prevent water from damaging the electronic components. Therefore, it is crucial to properly seal all interfaces between components. The MARESyE sensor is designed to handle these concerns, accommodate the optical requirements, and withstand 200 m depth ratings. The pressure at a given depth is directly proportional to the seawater density ($\rho_{sw} = 1025 \text{ kg m}^{-3}$), the acceleration of gravity, and the height of the water column. Thus, for 200 m depth, the housing must be able to withstand a 2 MPa hydrostatic pressure. To enable the integration of the sensor into any robotic platform, it is essential to minimize the weight and volume of the enclosure. A rectangular space-efficient design enables the optical configuration, and accommodates the internal processing within the same structure. However, the design's inherently flat surfaces make it more prone to pressure-induced buckling. To counteract this issue, the housing features rounded edges and a reinforcement structure, ensuring a highly durable enclosure. The use of AW6061 – T6 aluminum alloy is common for maritime applications [280] due to its favorable mechanical properties such as 275 MPa yield strength, low density of 2700 kg m^{-3} , and high corrosion resistance. Additionally, MARESyE's housing is anodized for further protection against corrosion. The compact internal space along with the heavy load in the cameras and processing unit leads to a fast rise in the global temperature of the system. This is countered by maximizing heat dissipation through direct contact of the cameras/processing unit with the aluminum housing, which itself dissipates the heat to the water.

The validation of this design is first conducted in simulation to study the stress and strain across the entire housing. A conservative linear static analysis is executed by applying a uniform pressure of 3 MPa to all surfaces in contact with water, ensuring an over-design of the sensor as a safety margin. Boundary conditions are set assuming that the housing is fixed by two of its back threaded holes, and its mesh is refined in the areas where the stress gradient is more accentuated. The results of the Finite Element Analysis (FEA) are depicted by figures 3.5 and 3.6, in terms of stress and displacements, respectively. The critical stress point of 227 MPa occurs in MARESyE's

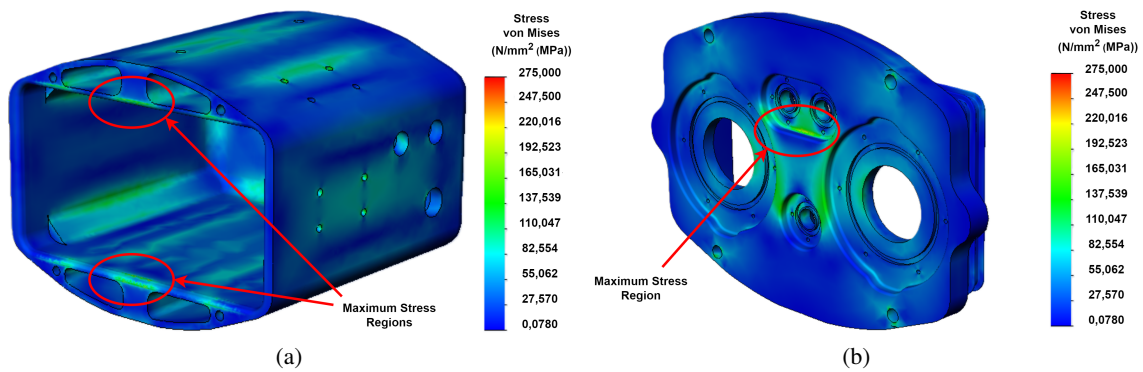


Figure 3.5: FEA stress results obtained from applying a uniform pressure of 3 MPa to the MARESyE housing structure. Figure a) depicts the maximum stress regions induced to the main housing body due to the contact with the lid's flange. Figure b) illustrates the maximum stress of 218 MPa endured by the housing lid's laser port edges.

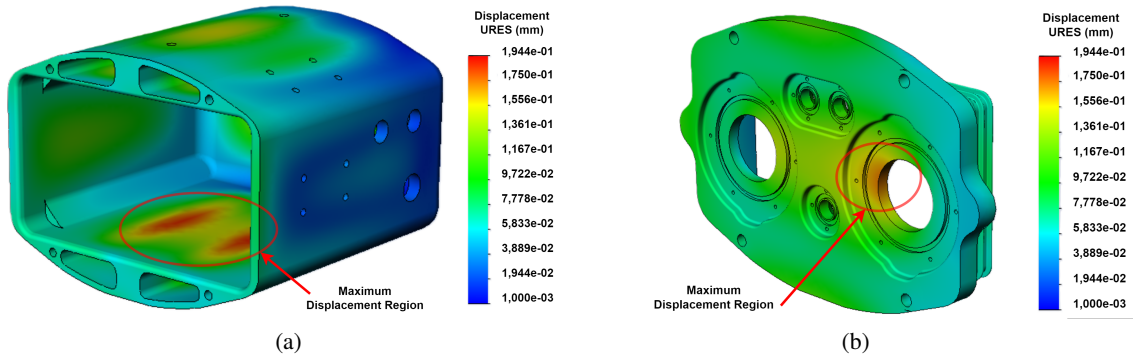


Figure 3.6: FEA displacement results obtained from applying a uniform pressure of 3 MPa to the MARESys housing structure. Figure a) depicts the maximum displacement of 1.9×10^{-4} m at the top and bottom walls of the main body housing due to its inherent large surface area. Figure b) illustrates the maximum displacement of 1.7×10^{-4} m near the center of the housing lid.

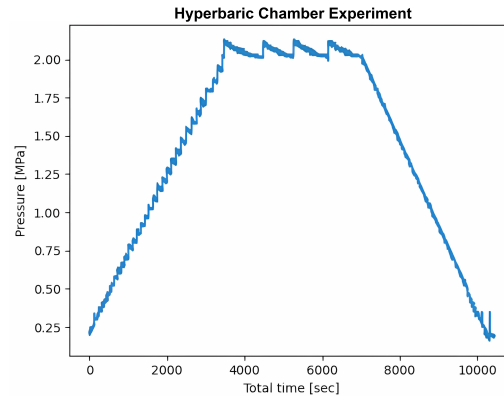


Figure 3.7: Response of the MARESys housing over time during the hyperbaric chamber experiment.

main housing body due to the pressure induced from the contact between the top and bottom walls with the lid's flange - figure 3.5a. A maximum displacement of 1.9×10^{-4} m deforms the center region of both top and bottom walls, as these are the largest surface areas - figure 3.6a. As for the lid, the critical stress point of 218 MPa can be found along the laser pocket edge, likely due to the abrupt geometric change and being the closest to the center - figure 3.5b. The peak displacement of 1.7×10^{-4} m is also registered near the center of the lid - figure 3.6b. The maximum stress experienced by both housing components remains well below the yield strength of the used aluminum alloy (275 MPa), thus the housing maintains its structural integrity when subjected to the 3 MPa hydrostatic pressures. A practical validation of MARESys's housing is later conducted in a hyperbaric chamber to simulate a pressure rating of 200 m. The experiment comprises three stages: a one-hour constant pressure increase until reaching 2 MPa; followed by a one-hour stage of stable 2 MPa pressure; and finally, a steady decrease over the course of the final hour. The response of the system is depicted in figure 3.7. The absence of abrupt pressure disruptions indicates a thorough sealing of the housing. Further assessment is conducted after

the chamber test to corroborate the water-tightness of the housing, as well as the lack of plastic deformations or fractures.

The mechanical design of the MARESyE sensor accomplishes a compact design (Height $\times$ Width $\times$ Depth) = (0.111 $\times$ 0.283 $\times$ 0.165) m that accommodates both the optical configuration and internal processing. Furthermore, the sensor is easy to integrate due to its low payload with underwater apparent weight of 0.210 kg, and total weight of 2.684 kg; while being capable of withstanding the hydrostatic pressures experienced at 200 m depth.

3.5 Results and Discussion

The content of this section intends to thoroughly analyze the performance of the introduced MARESyE hybrid imaging system. A comprehensive set of underwater experiments allow for a clear understanding of how the accuracy of the tri-dimensional measurements estimated by MARESyE varies according to the distance from the sensor. From an application perspective, MARESyE is employed to inspect and reconstruct sacrificial anode structures. Initially, experiments are conducted in a controlled underwater environment where measurements are collected from a fixed position. The sensor is later integrated within a commercial ROV and validated in real underwater environments at the ATLANTIS Test Center [6] where data collection is hindered by harsh conditions: heavy marine snow, constant lighting variations, and wave-induced motion blur.

3.5.1 Accuracy of the 3D Measurements

The distance at which the target is observed has direct impact on the quality of the retrieved tri-dimensional reconstruction. A myriad of factors contribute to this behavior, from the optical setup (stereo configuration baseline, camera parameterization, focal distance, image resolution) to environment inherent phenomena such as turbidity, light absorption and backscattering. This section aims to characterize the metric error in MARESyE's tri-dimensional readings, define a preferential operating range and validate its applicability for thorough inspection tasks where millimeter-accurate reconstructions are necessary. For this purpose, 3D measurements are collected from PS and LSR techniques while the MARESyE sensor is fixed relative to a 9 $\times$ 12 chessboard with 0.03 m sized squares. The aluminum target, with dimensions (Height $\times$ Width $\times$ Depth) = (0.4 $\times$ 0.3 $\times$ 0.06) m - refer to figure 3.8a - guarantees a flat surface plane from which consistent distance readings can be obtained from. The distance between MARESyE and the target is thoroughly measured using Leica DISTO X3² IR-based measuring device with reading resolution of $\pm$ 0.5 mm. A set of 20 readings were collected for each experiment to account for uncertainties of the IR sensor. This device employs a time-of-flight methodology to estimate depth relative to its closest straight-line reflection. As the target chessboard is mounted vertically, parallel to both MARESyE and the IR device, the distance obtained from the time-of-flight sensor is representative

²<https://shop.leica-geosystems.com/global/measurement-tools/disto/leica-disto-x3>

Table 3.3: SGBM parameters used to estimate the PS point clouds for every experiment.

SGBM Parameter	Value	SGBM Parameter	Value
minDisparity	0	P2	10000
numDisparities	120	disp12MaxDiff	5
blockSize	7	speckleWindowSize	150
P1	2300	speckleRange	10

of the whole flat target plane. For consistency, the camera parameters (brightness, gain, exposure, whitebalance) are fixed for every acquisition. Table 3.3 shows the SGBM parameters used during PS estimation throughout every experiment.

The subsequent metric analysis is conducted with basis on point cloud information gathered at 10 Hz over a span of 5 s, accounting to 50 samples per set. The following quantitative metrics are used along the remainder of the section to evaluate the estimated 3D readings:

- **MAE** - Mean Absolute Error:

$$MAE = \frac{\sum_{pc}^{n_s} |\hat{i}_{pc} - i|}{n_s}. \quad (3.9)$$

- **STD_{MAE}** - Standard Deviation of the MAE.

- **MRE** - Mean Relative Error:

$$MRE = \frac{\sum_{pc}^{n_s} \frac{MAE_{pc}}{i}}{n_s} \cdot 100, \quad (3.10)$$

where $\hat{i}_{pc}$ represents the average depth estimation for point cloud pc ; i is the average depth ground truth measurement; n_s corresponds to the number of point cloud samples.

Global Depth Characterization

The goal of the initial experimental setup is to describe the global accuracy of MARESyE's tri-dimensional measurements, i.e. how precise the sensor is in measuring distances in relation to itself. A set of stationary samples (PS and LSR readings, illustrated in figure 3.8b) is gathered at increasingly longer distances. Ground truth values are obtained by taking the distance from the face of the sensor to the chessboard for each set of measurements. The IR measuring device is fixed on top of the MARESyE sensor so that the variable distance - d_v - is thoroughly adjusted in 0.05 m increments, as depicted in figure 3.8c. The alignment between MARESyE and the IR sensor is ensured through a Computer Numerical Control (CNC) manufactured structural piece (tolerance of ± 0.1 mm) that serves as mount coupling the two sensors via the threaded holes in the imaging sensor's main housing. As a result, the transformation matrix between the two reference systems is known and remains consistent throughout every data acquisition step. This assembly is repeated for each experiment, as ground truth measurements are collected outside of water. As the ad-hoc underwater calibration procedure accurately maps the distortions caused by both the environment and the optical setup, the image formation plane is assumed to be approximately located at the camera sensor. The uncertainty introduced by this assumption is shown in table 3.2, and will

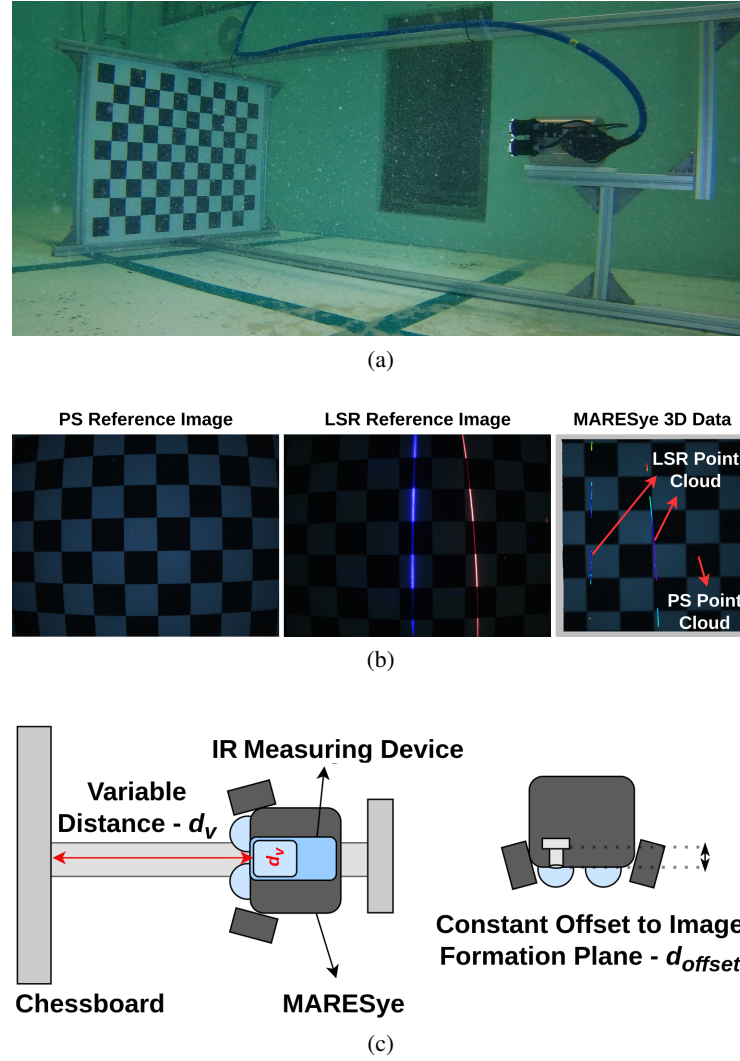


Figure 3.8: First experimental setup: assess the overall accuracy of MARESyE's 3D measurements by collecting PS and LSR point clouds at increasingly larger distances from a known target surface.

be further discussed in this section. An offset value, d_{offset} , is added as compensation to ensure a correct extrinsic calibration between the IR device and MARESyE. This offset coefficient is measured directly from MARESyE's Computer-Aided Design (CAD) model, which accurately reflects the real housing (CNC manufactured with ± 0.1 mm tolerance). The metric distance relative to MARESyE is estimated from each point cloud by averaging the depth values of all points belonging to the target surface plane.

The evolution of the mean absolute error and corresponding standard deviation is depicted in figure 3.9 for both PS and LSR tri-dimensional measurements. An exponential growth can be observed in the absolute error in the PS readings, as the distance from the sensor increases. The small baseline of the stereo configuration is adequate for close-range inspections - this is supported by the very precise depth estimations obtained along the $[0.3, 0.5]$ m distance range, achieving 0.0033 m of *MAE* and 0.87% *MRE* at most - table 3.4. For longer distances stereo-based systems

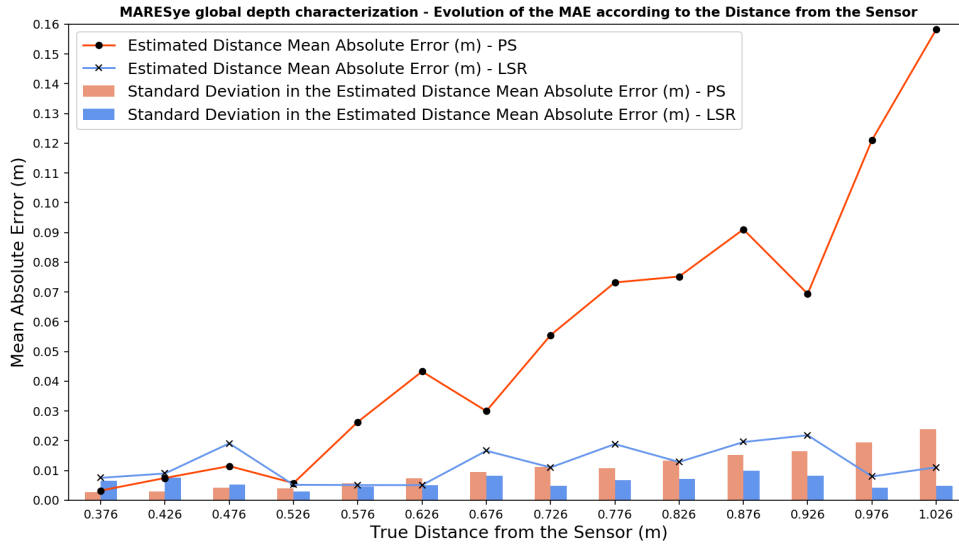


Figure 3.9: Evolution of the mean absolute error in the PS and LSR global depth estimations according to the distance from the MARESys sensor.

commonly lose resolution as the pixel matching algorithm progressively struggles to compute disparity values, hence the gradual increase in the MAE of the PS estimation, as well as its point cloud variability in terms of STD_{MAE} . MARESys employs two distinct laser projections for the LSR imaging methodology - red and blue beams. For this experimental setup both individual LSR estimations are aggregated and studied as a single one. The LSR depth estimation proves capable of maintaining a millimetric accuracy for most of the $[0.3, 0.6]$ m close-range spectrum, achieving at most 0.0051 m MAE and 0.89% MRE - refer to table 3.4. Even though the LSR estimations excel for measurements in close proximity, the worst case scenario sees LSR introduce a 0.0219 m MAE , corresponding to a 2.36% error relative to a 0.926 m ground truth reading. As the distance from the target increases, light absorption becomes noticeable in each of the laser projections. As consequence, segments of the beam become completely absent or faint of color, depending on how reflective the surface is. This can lead to a decrease in the number of triangulated points; on the other hand, these points become less variable since the more uniform color eases the segmentation process.

Table 3.4: MARESys global depth characterization - PS and LSR measurements.

True Dist. (m)	Photogrammetric Stereo					Light Stripe Ranging				
	Mean Estimated Dist. (m)	Estimated Dist. MAE (m)	Estimated Dist. STD_{MAE} (m)	Estimated Dist. MRE (%)	Mean Nr Points	Mean Estimated Dist. (m)	Estimated Dist. MAE (m)	Estimated Dist. STD_{MAE} (m)	Estimated Dist. MRE (%)	Mean Nr Points
0.376	0.3728	0.0033	0.0028	0.87%	62979	0.3685	0.0075	0.0066	2.01%	1000
0.426	0.4186	0.0075	0.0030	1.76%	62802	0.4170	0.0090	0.0076	2.11%	856
0.476	0.4645	0.0115	0.0043	2.42%	63597	0.4569	0.0191	0.0052	4.02%	1033
0.526	0.5203	0.0058	0.0041	1.10%	63873	0.5208	0.0052	0.0030	0.99%	753
0.576	0.5497	0.0263	0.0057	4.57%	63819	0.5709	0.0051	0.0047	0.89%	772
0.626	0.5828	0.0433	0.0074	6.91%	64125	0.6209	0.0051	0.0050	0.81%	754
0.676	0.6461	0.0300	0.0096	4.43%	63527	0.6594	0.0166	0.0083	2.46%	807
0.726	0.6705	0.0555	0.0112	7.64%	63865	0.7150	0.0110	0.0048	1.52%	932
0.776	0.7028	0.0732	0.0109	9.43%	63790	0.7572	0.0189	0.0068	2.43%	903
0.826	0.7509	0.0752	0.0132	9.10%	62624	0.8132	0.0129	0.0073	1.56%	1052
0.876	0.7850	0.0911	0.0151	10.39%	59775	0.8565	0.0196	0.0100	2.24%	976
0.926	0.8567	0.0694	0.0166	7.49%	62935	0.9042	0.0219	0.0083	2.36%	871
0.976	0.8550	0.1210	0.0194	12.40%	61514	0.9680	0.0080	0.0041	0.82%	928
1.026	0.8678	0.1582	0.0239	15.42%	61585	1.0150	0.0111	0.0049	1.08%	920

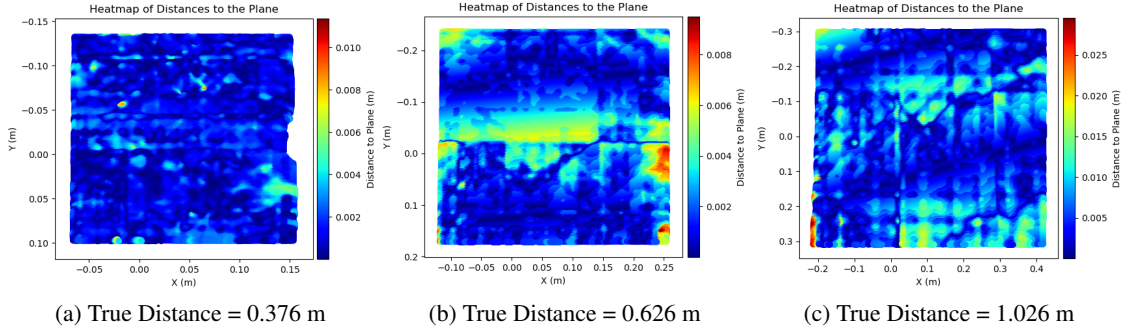


Figure 3.10: A study on the flatness of MARESyE's PS point clouds was conducted for three distinct metric ranges. The RANSAC algorithm was used to approximate the 3D chessboard representations with a plane. The resulting heatmap of the distances from each point to this plane shows an arbitrary spatial distribution. This further validates the underwater calibration process, as the measurements do not indicate systematic errors from faulty distortion mapping.

The uncertainty induced by the stereo calibration error, presented in table 3.2, is sub-millimeter for the $[0.3, 0.876]$ m range. When ΔZ reaches a millimeter, for distances over 0.876 m, the estimated MAE for MARESyE's PS point clouds is circa 0.090 m. In essence, the ΔZ uncertainty is always an order of magnitude below the estimated MAE error, thus being negligible for the conclusions presented along this characterization study. Since the 3D error characterization presented along this work relies on the mean value of the Z coordinate for each point cloud, any distortions not properly corrected by calibration could introduce systematic errors into the measurements. To study this behavior, the flatness of the tri-dimensional PS representation of the target chessboard is approximated by a plane using the RANDOM SAMPLE Consensus (RANSAC) algorithm [281]. The distances from each point to the estimated flat plane are calculated using equation 3.11:

$$d_{RANSAC} = \frac{|a_{plane}P_X + b_{plane}P_Y - P_Z + c_{plane}|}{\sqrt{a_{plane}^2 + b_{plane}^2 + 1}}, \quad (3.11)$$

where (P_X, P_Y, P_Z) represents a point in the point cloud; and $a_{plane}, b_{plane}, c_{plane}$ are the plane coefficients. Figure 3.10 presents the distance heatmap for three case studies representing close, medium, and long-range experiments. This visual representation indicates that the error is spatially distributed in each of the point clouds, regardless of the distance to the observer, and not contained in the periphery of the image as it is common for optical systems affected by pincushion and barrel distortion effects, or optical misalignments induced by the dome [278]. MARESyE's 3D estimations do not exhibit bias or uncertainty resulting from the underwater sensor calibration, as the applied procedure effectively mitigates both tangential and radial distortions.

The MARESyE system achieves the best performance within the $[0.3, 0.55]$ m range of operation, with both the PS and LSR methodologies capable of computing point cloud with millimetric accuracy. As the distance from the target increases, the error in the PS point cloud rises exponentially due to the short baseline length. The LSR estimations maintain an average *MRE* of 1.81% over the whole range spectrum. The slight increase in error, noticeable for higher distances, can

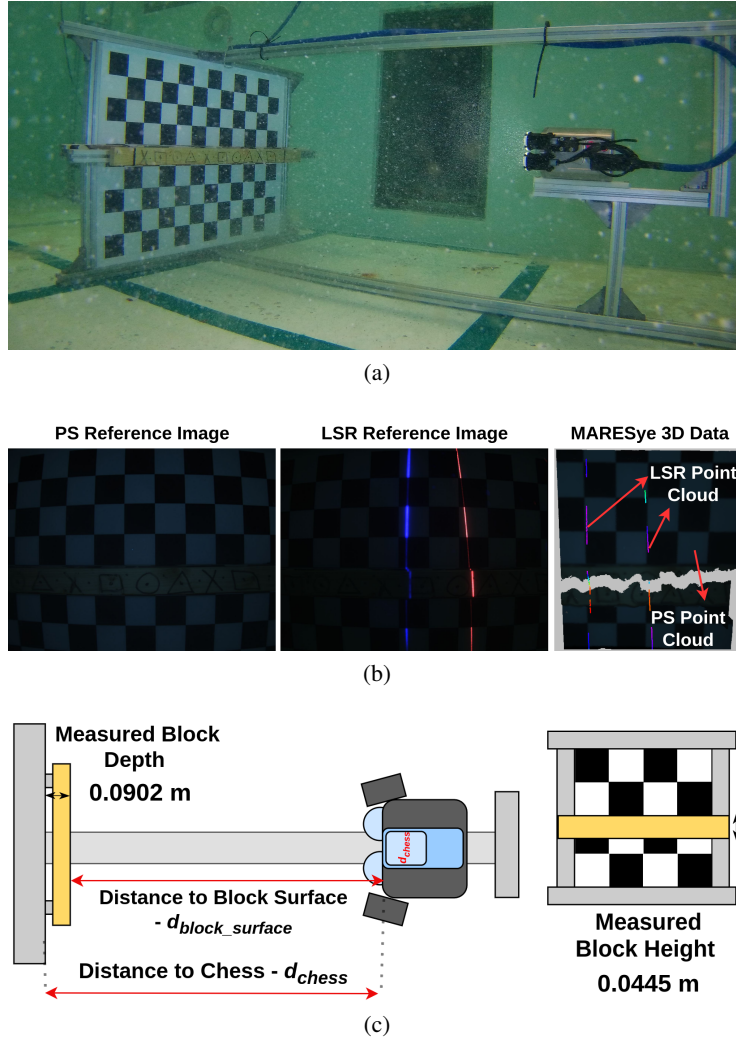


Figure 3.11: Second experimental setup: analyze MARESyE’s PS and LSR relative depth estimation accuracy for increasingly larger distances from the sensor.

be justified by the segmentation difficulties introduced by the underwater absorption phenomenon that sees the beam projections get progressively more faint and difficult to segment. This analysis supports the benefit of fusing both imaging techniques to improve the overall performance of the MARESyE sensor, maintaining a dense and textured point cloud and increasing its range of operation [38].

Relative Depth Characterization

The second set of experiments aims to examine how the accuracy of relative depth measurements estimated by the MARESyE system evolves with respect to the distance from the target. Figure 3.11 illustrates the setup for this study. A 0.0902 m thick block is placed in front of the chessboard, and a set of stationary samples from both imaging techniques is acquired (figure 3.11b). The thickness of the block is estimated from each point cloud by measuring the global

depth difference between both surface planes - refer to equation 3.12. The mean standard deviation of this measurement ($STD_{MAE_{block_depth}}$) is given by equation 3.13. Camera parameterization is maintained in between experiments.

$$MAE_{block_depth} = MAE_{d_{chess}} - MAE_{d_{block}}, \quad (3.12)$$

$$STD_{MAE_{block_depth}} = \sqrt{STD_{MAE_{d_{chess}}}^2 + STD_{MAE_{d_{block}}}^2}. \quad (3.13)$$

In contrast to the global depth characterization experiments, the PS proves to be the more stable method when estimating relative depth. Figure 3.12 depicts the head-to-head comparison for the evolution of the MAE_{block_depth} in the estimation of both the PS and LSR algorithms. The stereo-based methodology maintains the same behavior of increasingly higher $STD_{MAE_{block_depth}}$. On the other hand, the MAE_{block_depth} is contained within a millimeter for the [0.5, 0.6] m operation range. For distances farther away from the sensor, the mean absolute error slightly increases and stabilizes around the 0.01 m mark - refer to table 3.5 - corresponding to a MRE of 12.8% for the studied distance spectrum. As the distance from the sensor increases, the PS imaging technique struggles to compute pixel correspondences leading to the previously demonstrated exponential curve in the PS global depth estimation (figure 3.9). This behavior directly impacts the relative depth readings, since the algorithm progressively calculates similar disparity values regardless of the distance to the sensor, inducing a saturation in the MAE_{block_depth} .

The relative measurements estimated through the LSR technique achieve at most 0.004 m of MAE_{block_depth} for close-range readings, maintaining this millimetric accuracy in the [0.5, 0.6] m spectrum - table 3.6. Outside of this range, the mean absolute error progressively increases with distance, reaching saturation for distances above 0.9 m. The reason for this drastic decay in performance over distance is the extrinsic laser-camera configuration. The small angle deviation

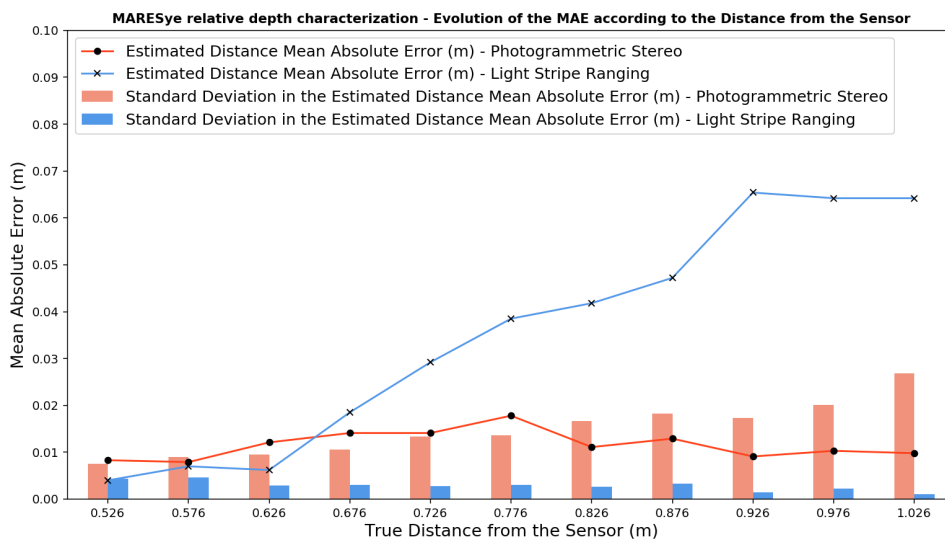


Figure 3.12: Evolution of the mean absolute error in the PS and LSR relative depth estimations according to the distance from the MARESy sensor.

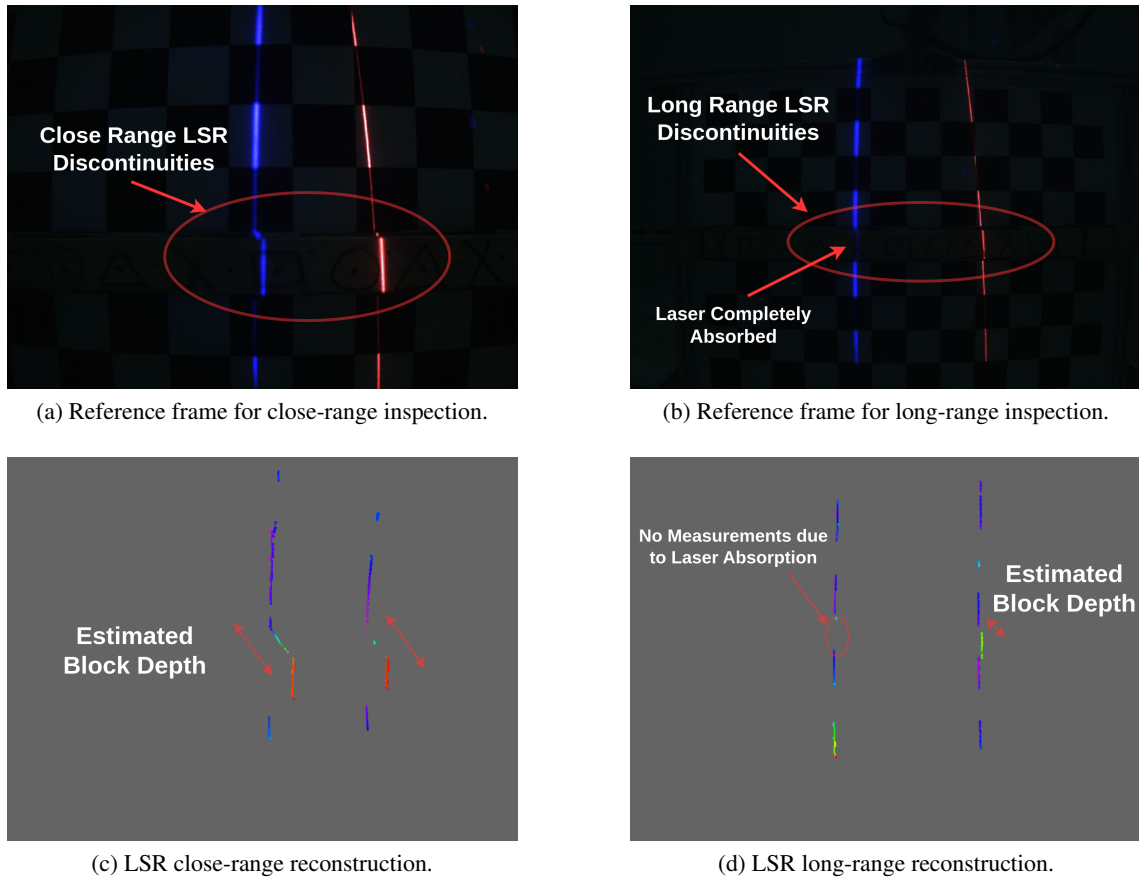


Figure 3.13: Impact of the distance from the MARESyE sensor in the LSR relative depth estimation. Close-range inspections create noticeable discontinuities that are leveraged to calculate an accurate block depth. In addition to suffering from laser beam absorption, LSR long-range reconstructions struggle progressively due to lack of resolution in the laser-camera setup. This leads to increasingly smaller discontinuities, and consequently less accurate relative depth estimations.

between both projection planes - a compromise enforced to secure the synergy between optical setup and internal housing dimensions - constrains the resolution of the LSR method. A clear example of this behavior is illustrated in figure 3.13. For close-range inspections such as figure 3.13a, there is a noticeable discontinuity in the laser projections that clearly separate the beam segments that are projected onto the block from the ones reflected back from the chessboard, leading to a straightforward point cloud triangulation - refer to figure 3.13c. The same cannot be said for longer range reconstructions. Apart from the aforementioned light absorption hindrances, the discontinuities in the beam projections become smaller and smaller, making it very hard for the LSR algorithm to estimate different depth values for each of the beam projection planes, as seen in figure 3.13d.

In essence, the MARESyE sensor performs best within the $[0.5, 0.6]$ m range, with both the PS and LSR imaging techniques achieving millimetric accuracy for the relative depth estimation task. The LSR methodology retrieves extremely accurate point clouds in this close-range spectrum,

Table 3.5: MARESy relative depth and spatial dimensions characterization - photogrammetric stereo measurements.

True Dist. (m)	Relative Depth Estimation Analysis				Spatial Dimensions Estimation Analysis				Mean Nr Chess Points	Mean Nr Block Points
	Mean Estimated Block Depth (m)	Estimated Block Depth MAE (m)	Estimated Block Depth STD _{MAE} (m)	Estimated Block Depth MRE (%)	Mean Estimated Block Height (m)	Estimated Block Height MAE (m)	Estimated Block Height STD _{MAE} (m)	Estimated Block Height MRE (%)		
0.526	0.0985	0.0083	0.0076	9.25%	0.0497	0.0052	0.0005	11.60%	53713	6218
0.576	0.0981	0.0079	0.0090	8.74%	0.0461	0.0016	0.0006	3.49%	52639	6913
0.626	0.1023	0.0121	0.0095	13.42%	0.0443	0.0002	0.0000	0.39%	55254	5405
0.676	0.1043	0.0141	0.0106	15.65%	0.0457	0.0012	0.0007	2.79%	50444	4907
0.726	0.1043	0.0141	0.0133	15.67%	0.0471	0.0026	0.0010	5.75%	52710	4710
0.776	0.1080	0.0178	0.0136	19.78%	0.0439	0.0006	0.0013	1.44%	50854	4205
0.826	0.1013	0.0111	0.0166	12.26%	0.0472	0.0027	0.0011	6.01%	48358	4065
0.876	0.1031	0.0129	0.0182	14.33%	0.0543	0.0098	0.0007	22.12%	47499	3657
0.926	0.0993	0.0091	0.0173	10.10%	0.0530	0.0085	0.0011	19.21%	44133	3068
0.976	0.1005	0.0103	0.0201	11.46%	0.0555	0.0110	0.0014	24.66%	39864	2966
1.026	0.1000	0.0098	0.0268	10.85%	0.0554	0.0109	0.0020	24.43%	47861	2596
True Block Depth - 0.0902 m					True Block Height - 0.0445 (m)					

Table 3.6: MARESy relative depth and spatial dimensions characterization - light stripe ranging 3D measurements.

True Dist. (m)	Relative Depth Estimation Analysis				Spatial Dimensions Estimation Analysis				Mean Nr Chess Points	Mean Nr Block Points
	Mean Estimated Block Depth (m)	Estimated Block Depth MAE (m)	Estimated Block Depth STD _{MAE} (m)	Estimated Block Depth MRE (%)	Mean Estimated Block Height (m)	Estimated Block Height MAE (m)	Estimated Block Height STD _{MAE} (m)	Estimated Block Height MRE (%)		
0.526	0.0942	0.0040	0.0043	4.43%	0.0517	0.0072	0.0001	16.24%	392	280
0.576	0.0972	0.0070	0.0046	7.77%	0.0464	0.0030	0.0012	6.68%	354	238
0.626	0.0840	0.0062	0.0029	6.87%	0.0471	0.0026	0.0013	5.77%	341	185
0.676	0.0717	0.0185	0.0030	20.49%	0.0434	0.0022	0.0003	4.87%	326	190
0.726	0.0610	0.0292	0.0028	32.41%	0.0452	0.0009	0.0003	1.93%	287	192
0.776	0.0517	0.0385	0.0031	42.72%	0.0415	0.0041	0.0006	9.28%	444	166
0.826	0.0484	0.0418	0.0027	46.34%	0.0391	0.0054	0.0004	12.20%	421	192
0.876	0.0430	0.0472	0.0033	52.29%	0.0385	0.0060	0.0093	13.52%	518	177
0.926	0.0248	0.0654	0.0015	72.53%	0.0377	0.0068	0.0001	15.31%	241	85
0.976	0.0260	0.0642	0.0023	71.15%	0.0328	0.0117	0.0001	26.32%	249	77
1.026	0.0260	0.0642	0.0011	71.18%	0.0312	0.0133	0.0002	29.97%	158	61
True Block Depth - 0.0902 m					True Block Height - 0.0445 (m)					

which can once more be leveraged to guide the dense and textured PS estimation. As the distance from the sensor increases, the MARESy system should progressively attribute more confidence to the PS readings as these maintain performance for long-range estimations.

Spatial Dimensions Characterization

From the second set of experiments presented above, the MARESy sensor's performance can also be evaluated when estimating spatial dimensions. The height of the block assembled in the setup presented in figure 3.11 is now the target study. For each set of PS and LSR point clouds, a bounding box is estimated from the points that belong to the block surface. The estimated height of this bounding box is then compared to the measured 0.0445 m ground truth value. The obtained results are presented in table 3.5 and 3.6 for the PS and LSR imaging techniques, respectively. Moreover, figure 3.14 depicts the height MAE evolution according to the distance from MARESy, for each methodology. The two algorithms maintain millimetric accuracy for most of the studied spectrum. The long distance performance decay is seen once again with both PS and LSR tending to MAE around the 0.01 m in height. Nevertheless, this occurs due to distinct circumstances: the stereo-based point clouds estimate progressively higher height readings since the poor visibility induces pixel matching errors in the chessboard/block transitional region; on the other hand, the triangulated LSR points tend to shrink the calculated height due to the absorption of portions of the beam segment as well as the continuous decrease in the laser projection discontinuities that induce faulty estimations.

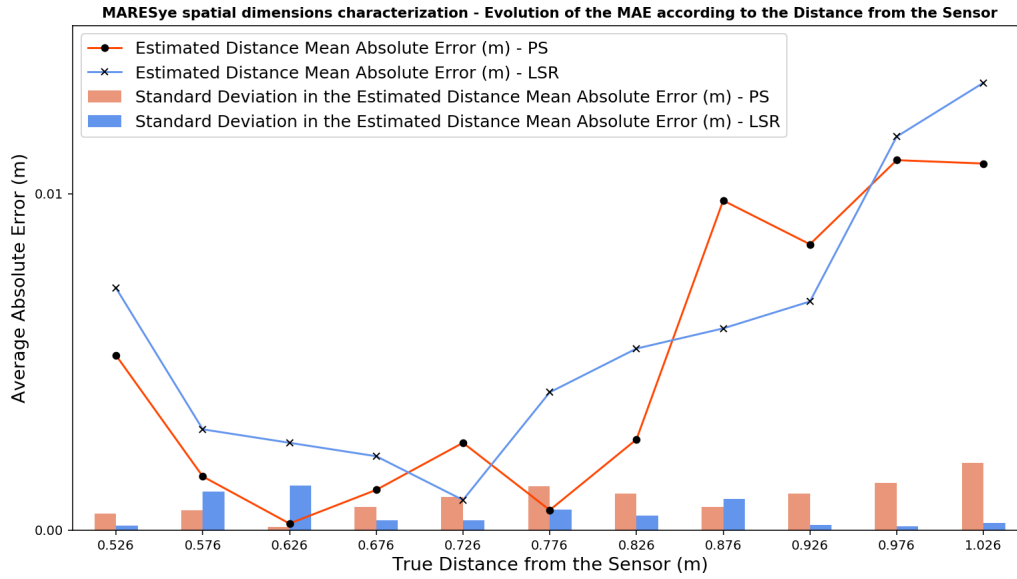


Figure 3.14: Evolution of the mean absolute error in the PS and LSR spatial dimension estimation according to the distance from the MARESyE sensor.

3.5.2 Inspecting Submerged Sacrificial Anodes

The maritime environments that offshore platforms are exposed to are extremely harsh. Sacrificial anodes are used to prevent the rapid degradation of these structures. The goal is to extend the platform's lifespan and guarantee its structural integrity by having a sacrificial layer corrode in place of the steel structure. As a result, periodic maintenance and replacement are necessary to ensure continued protection against corrosion. These operations are extremely complicated due to the environmental conditions, and commonly resort to divers with handheld devices where the information is later processed. The use of underwater imaging systems capable of acquiring high-fidelity 2D and 3D information can lead to the automation of these O&M tasks, while guaranteeing a reliable and repeatable analysis of the anode condition. This section employs the MARESyE sensor to reconstruct real sacrificial anodes, while scrutinizing its performance: first within a controlled underwater environment, and later by integrating the sensor in a commercial ROV that is deployed in a real underwater application at the ATLANTIS Test site.

Controlled Underwater Experiments

In the initial application experiment, MARESyE is employed to reconstruct a real decommissioned sacrificial anode within an indoor pool - refer to figure 3.15. This anode was cut off from the DURIUS floating structure [41] after a long time-span of operation, leading to an extremely irregular surface. Due to the absence of ground truth data or information regarding the original dimensions of the anode, relative measurements were collected to serve as baseline for comparison against the 3D readings collected from the sensor. The upper and lower portions of the structure are analyzed independently. Measurements corresponding to the peaks (most salient regions of the surface), and valleys (innermost eroded regions) were thoroughly noted; the distance to the metal

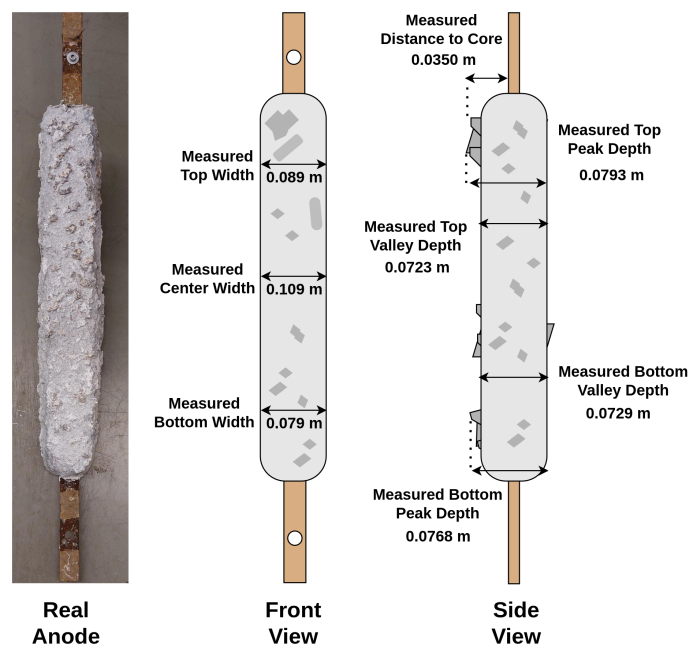


Figure 3.15: Decommissioned sacrificial anode cut off from the DURIUS floating structure [41] used during the controlled underwater experiments. Relative depth measurements are thoroughly collected from its irregular surface and used as ground truth values to evaluate MARESyE's performance within real applications.

core of the anode is measured from each peak to the corresponding extremity; and finally the width of the anode is studied along its different sections - figure 3.15. The relative depth characteristics are obtained by using the Leica DISTO X3 IR sensor to measure the distance between each control feature and the background wall. Spatial-based dimensions are measured with a digital caliper³ with reading accuracy of up to ± 0.02 mm. These ground truth measurements are repeated 20 times for both spatial and relative characteristics to minimize uncertainty and ensure consistent, reliable metric readings. The experimental setup secures the anode in a fixed position against the wall of the pool, with MARESyE located approximately 0.55 m away, facing the structure. Relative measurements using the stereo-based methodology are derived from the depth values between the anode surface and the background pool wall. The LSR acquisitions guarantee that one laser reflects off the anode structure, while the other is projected onto the wall, facilitating measurements between both triangulated point clouds to represent relative depth. The spatial analysis of each point cloud section is performed by selecting the two outermost leveled points on either side of the anode's surface and calculating the metric distance between them. This picking methodology is enabled by the sub-millimeter spatial resolution of the PS point cloud (see table 3.1): within the data collection range, MARESyE estimates point clouds with a spatial resolution of 0.00033 m. Measurements are repeated for each sample, collected at 10 Hz over a 5 s interval.

The analysis of relative depth from 3D data obtained through both PS and LSR imaging techniques is detailed in table 3.7, while figure 3.16 visually presents the captured reference frames

³<https://docs.rs-online.com/2437/A700000009024162.pdf>

Table 3.7: Analysis of MARESyE’s relative depth measurements in the 3D reconstruction of a sacrificial anode within a controlled underwater environment. Measurements are estimated by averaging the readings from each point cloud sample.

Characteristics		Ground Truth	Photogrammetric Stereo			Light Stripe Ranging		
		Measured (m)	Estimated (m)	Absolute Error (m)	Relative Error (%)	Estimated (m)	Absolute Error (m)	Relative Error (%)
Top	Peak Depth	0.07930	0.08101	0.00171	2.16%	0.07311	0.00619	7.81%
	Valley Depth	0.07230	0.07447	0.00217	3.00%	0.06537	0.00693	9.59%
	Distance to Core	0.03500	0.03897	0.00397	11.34%	0.04275	0.00775	22.14%
	Anode Wear	0.00700	0.00654	0.00046	6.57%	0.00774	0.00074	10.57%
Bottom	Peak Depth	0.07680	0.07771	0.00091	1.18%	0.07200	0.00480	6.25%
	Valley Depth	0.07290	0.07370	0.00080	1.10%	0.06527	0.00763	10.47%
	Distance to Core	0.03500	0.04170	0.00670	19.14%	0.04624	0.01124	32.11%
	Anode Wear	0.00390	0.00401	0.00011	2.82%	0.00673	0.00283	72.56%

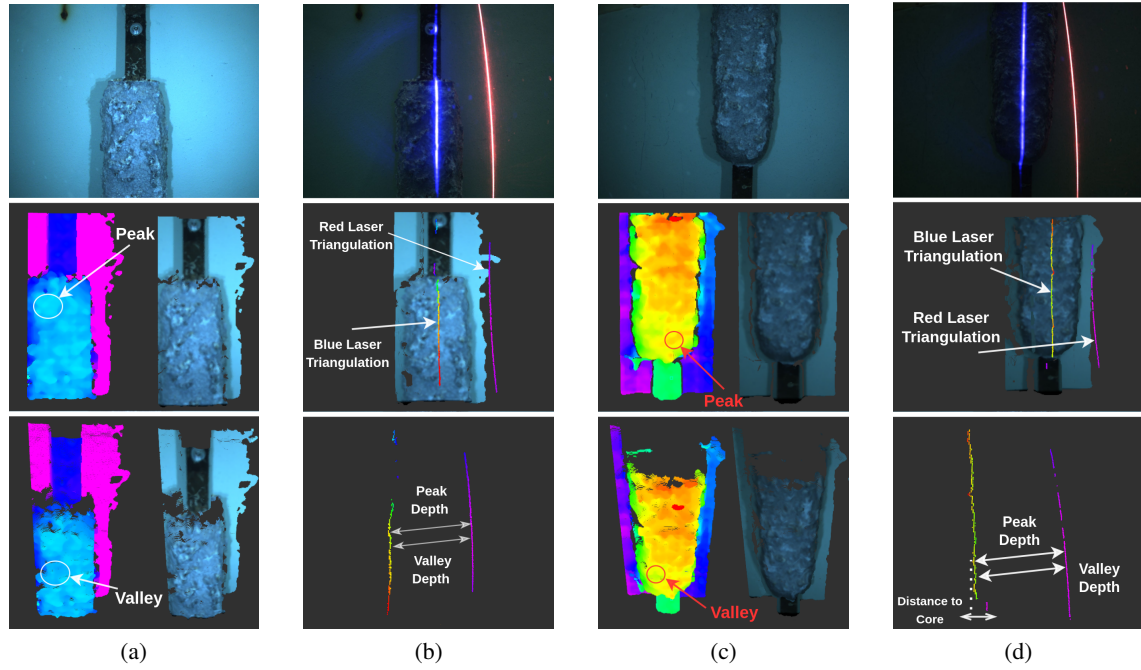


Figure 3.16: MARESyE 3D reconstructions within a controlled indoor pool scenario: photogrammetric stereo (a, c) and light stripe ranging (b, d) point clouds and corresponding reference frames for the top and bottom sections of the sacrificial anode.

and their corresponding tri-dimensional reconstructions. The photogrammetric stereo algorithm excels in scenarios characterized by low turbidity and abundant texture. The estimated PS 3D reconstructions, illustrated in figure 3.16a and 3.16c, intricately capture the details of the eroded anode, providing millimetric precision for its peaks and valleys. For the top section of the structure, the depth values for peaks and valleys are estimated at 0.08101 m and 0.07447 m, respectively. This corresponds to absolute errors of 0.00171 m and 0.00217 m. Anode wear, defined as the depth difference between peaks and valleys, is calculated by the stereo-based method with an absolute error of 0.00046 m. Enhanced performance is observed in the experiments pertaining to the bottom part of the anode, where the algorithm estimates wear at 0.00401 m with an absolute error

Table 3.8: Analysis of MARESys's photogrammetric stereo spatial dimensions measurements in the 3D reconstruction of a sacrificial anode within a controlled underwater environment. Measurements are estimated by averaging the readings from each point cloud sample.

Characteristics	Ground Truth	Photogrammetric Stereo		
	Measured (m)	Estimated (m)	Absolute Error (m)	Relative Error (%)
Top Width	0.0890	0.0907	0.0017	1.96%
Center Width	0.1090	0.1087	0.0003	0.31%
Bottom Width	0.0790	0.0778	0.0012	1.54%

of 0.00011 m. These measurements are derived from very precise peak and valley readings that translate 1.18% and 1.10% relative error in the measurements, respectively. The light stripe ranging technique also achieves millimetric accuracy in its measurements, although its performance is somewhat inferior when compared to the previously mentioned stereo-based reconstructions. In figures 3.16b and 3.16d, a slight overexposure is noticeable in both laser line projections. This overexposure can introduce enough variability in the triangulated point clouds to influence the readings, as LSR measurements have a limited number of points representing peaks and valleys compared to the photogrammetric stereo point cloud. Nevertheless, LSR accurately measures the depth of the anode's top peak and valley with absolute errors of 0.00619 m and 0.00693 m, respectively. The calculated anode wear is 0.00774 m, resulting in a 10.57% error relative to the ground truth reading of 0.007 m. Each section of the anode is linked to an extremity serving as a mounting support for the structure. As these surfaces are directly connected to the metal core of the anode, the distance from peak to extremity surface is defined as the distance to the core. The predominantly dark and uniform texture of this region hinders the PS pixel correspondence operation, resulting in an absolute error 0.00397 m for the top section and 0.00670 m for the lower portion. The higher deviation in the measurement for the bottom half of the anode can be attributed to the darker illumination conditions, given its placement further down from the water surface. The LSR method overshoots both measurements, primarily due to the aforementioned issues of variability. These issues are aggravated in this estimation, as the beam segments projected onto the anode and extremities display considerable variations in pixel intensity. Shifting the focus to a spatial dimensions analysis, the photogrammetric stereo algorithm demonstrates reliability by accurately estimating the anode's width across its entire surface. The metric results, as presented in table 3.8, align with the findings illustrated in figure 3.14, where millimetric accuracy is achieved for close-range measurements. In the worst-case scenario, the calculated width for the upper half of the structure exhibits an absolute error of 0.0017 m. For the center and bottom sections, the stereo-based approach estimates a width of 0.1087 m and 0.0778 m, resulting in relative errors of 0.31% and 1.54%, respectively. In summary, the MARESys sensor is capable of achieving the millimetric precision essential during inspection operations. Both imaging techniques prove reliable in a controlled environment with clear visibility and no external hindrances such as the constant lighting changes, heavy marine snow and motion blur encountered during real operations. The following section sees MARESys integrated in a commercial ROV, and deployed in a real maritime

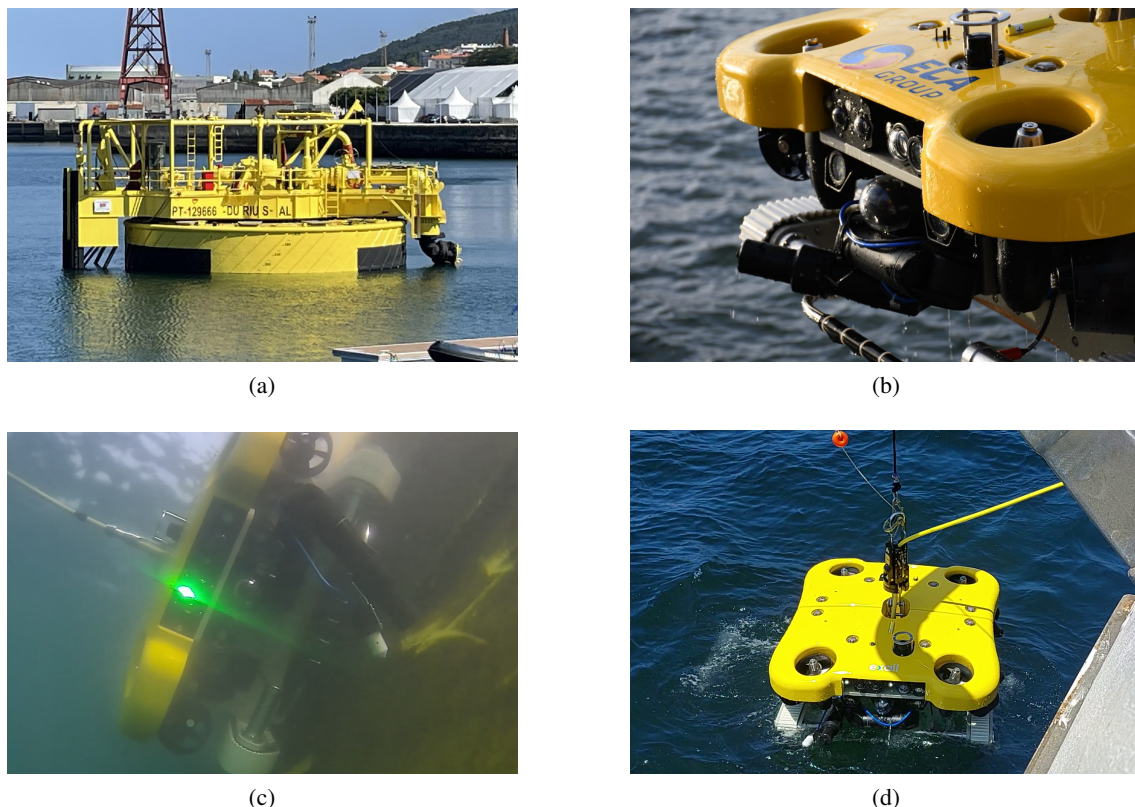


Figure 3.17: Set of underwater experiments conducted at the ATLANTIS Coastal Test Center. Figure a) illustrates DURIUS, the floating platform that allows for near-real test conditions. Figure b), c) and d) show the MARESyE sensor integrated in Exail's ROVINGBAT ROV, used during the inspection operation.

environment where the impact of these real-world conditions can be studied.

Real Underwater Application - ATLANTIS Coastal Testbed

The final series of experiments takes place at the ATLANTIS Coastal Test Center in Viana do Castelo, Portugal. This testbed provides access to DURIUS, a repurposed floating buoy originally used in the Oil & Gas industry. DURIUS serves as an ideal platform for testing and validating maritime technologies in a near-real environment [41]. Figure 3.17a illustrates the DURIUS structure, which is the focal point of inspection in this section. For the inspection operation at the ATLANTIS Coastal Test Center, the MARESyE sensor is integrated into Exail's ROVINGBAT ROV⁴ — refer to figure 3.17b. The ROVINGBAT vessel is capable of landing and securing itself on DURIUS through the use of four vertical thrusters. Additionally, it can navigate along the structure by employing two crawlers, ensuring stability for precise measurements during operations. To inspect and reconstruct the sacrificial nodes utilized to safeguard the DURIUS platform against corrosion, the ROVINGBAT, integrated with the MARESyE sensor, first navigates in close

⁴<https://www.ecagroup.com/en/solutions/rovingbat-hybrid-rov>

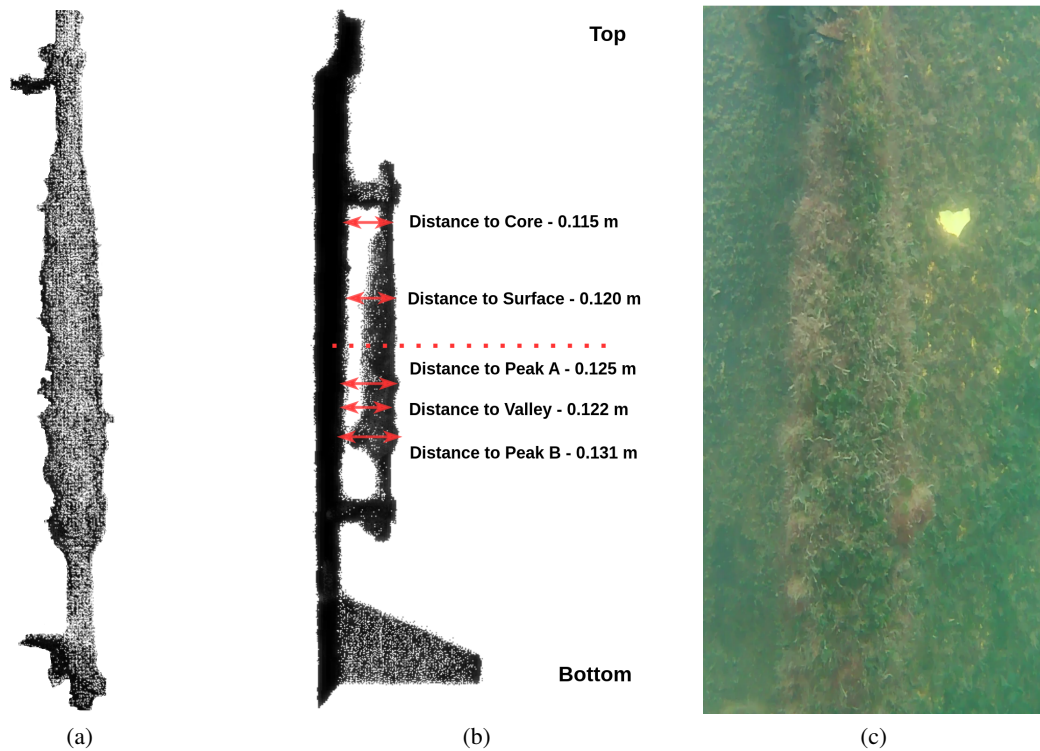


Figure 3.18: Sacrificial anode from the DURIUS structure used as the reconstruction target. Figures a) and b) illustrate the 3D model segmented from the dry dock reconstruction of the DURIUS platform and the respective characteristics used for analysis. Figure c) displays the state of the anode, covered in biofouling, at the time of inspection.

proximity of the buoy. The ROV then holds a stable position facing the target anode at approximately 0.60 m. MARESyE then estimates tri-dimensional measurements from both the PS and LSR techniques. These measurements are transmitted with negligible delay via a fiber optic tether that connects the ROV to shore, being subsequently used by the operator for a detailed inspection procedure.

Before being anchored at Viana do Castelo's Port, the DURIUS structure underwent rigorous maintenance and cleaning after a long time-span of offshore deployment. During this downtime, a thorough 3D reconstruction was conducted on land, resorting to a Leica BLK360⁵ laser scan system with accuracy of up to 4 mm. The complete reconstruction of the DURIUS buoy is obtained by stitching together each point cloud through a Multi-scale Point-to-point ICP algorithm [282]. For the scope of this discussion, one of the inspected sacrificial anodes is segmented from the generated 3D model - figure 3.18a. The highly eroded sacrificial anode leaves the core almost exposed on its top section. On the other hand, the bottom part shows a more irregular pattern with peaks and valleys that represent different levels of corrosion. Relative measurements, describing the depth readings between the highly irregular surface of the anode and the main body of the DURIUS platform are collected directly from the reconstructed mesh - figure 3.18b. These are

⁵<https://shop.leica-geosystems.com/leica-blk/blk360/technology>

Table 3.9: Analysis of MARESyE’s relative depth measurements in the 3D reconstruction of DURIUS’ sacrificial anode within a real underwater environment at the ATLANTIS Test Center.

Characteristics		Ground Truth	Photogrammetric Stereo		Light Stripe Ranging	
		Measured (m)	Estimated (m)	Absolute Error (m)	Estimated (m)	Absolute Error (m)
Top	Distance to Surface	0.120	0.1443	0.0243	0.1245	0.0045
	Distance to Core	0.115	0.1377	0.0227	0.1189	0.0039
Bottom	Distance to Peak A	0.125	0.1503	0.0253	0.1568	0.0318
	Distance to Valley	0.122	0.1463	0.0243	0.1319	0.0099
	Distance to Peak B	0.131	0.1606	0.0296	0.1682	0.0372

subsequently used to analyze the performance of the MARESyE sensor in this environment.

As expected, gathering tri-dimensional data in real environments opens up additional challenges when compared to the inspection conducted at the indoor pool: the ROVINGBAT’s thrusters stir up a lot of the dust and particles, vastly increasing the backscattering phenomena; strong waves that are capable of swaying the ROV’s hovering difficult the image acquisition process; constant shifts in light conditions lead to uneven lighting of the scene, and the absorption of projected laser beams hampers the triangulation process. MARESyE’s range of operation is also limited by the vehicle’s maximum depth rating of 100 m, and the length of the fiber optics umbilical cable (170 m). Additionally, the underwater inspection operation was conducted several weeks after the cleaning procedure. This results in a thick layer of biofouling (moss, mussels, etc) that completely covers the anode structure - refer to figure 3.18c.

The photogrammetric stereo algorithm provides a dense and detailed 3D estimation of the scenario, enabling the assessment of the anode’s characteristics, as illustrated in figure 3.19a. As per the data presented in table 3.9, a systematic average absolute error of 0.0251 m is observed across all the analyzed characteristics. Despite introducing a rich and textured environment that favors stereo-matching computation, the aggregation of materials is responsible for a large percentage of the skew in these measurements. Leveraging the findings in table 3.5, an MAE of 0.0121 m can be expected in the relative measurements estimated through the PS methodology, for the 0.60 m distance range where the tri-dimensional data is collected. Moreover, the table indicates a consistent overshooting tendency in the estimated relative depth by the stereo-based algorithm. This observation supports the conclusion that a portion of the absolute error in the anode’s reconstruction corresponds to this overshoot. The average biofouling thickness can thus be roughly calculated as approximately 0.0130 m, by taking the difference between the systematic 0.0251 m error and the expected MAE for the acquisition distance. The interpretation of the measurements estimated by the light stripe ranging method is not as straightforward. The evaluation of said readings is conducted in table 3.9. The presented relative measurements are collected by taking the difference in depth estimation from each of the beams: the blue laser reflects back from the anode; while the red beam is projected onto the DURIUS structure - refer to figure 3.19b. As per the experiments presented in figure 3.9, the global depth estimation error in the LSR point cloud is mostly constant across the range spectrum where data was acquired, therefore any error does not affect the relative

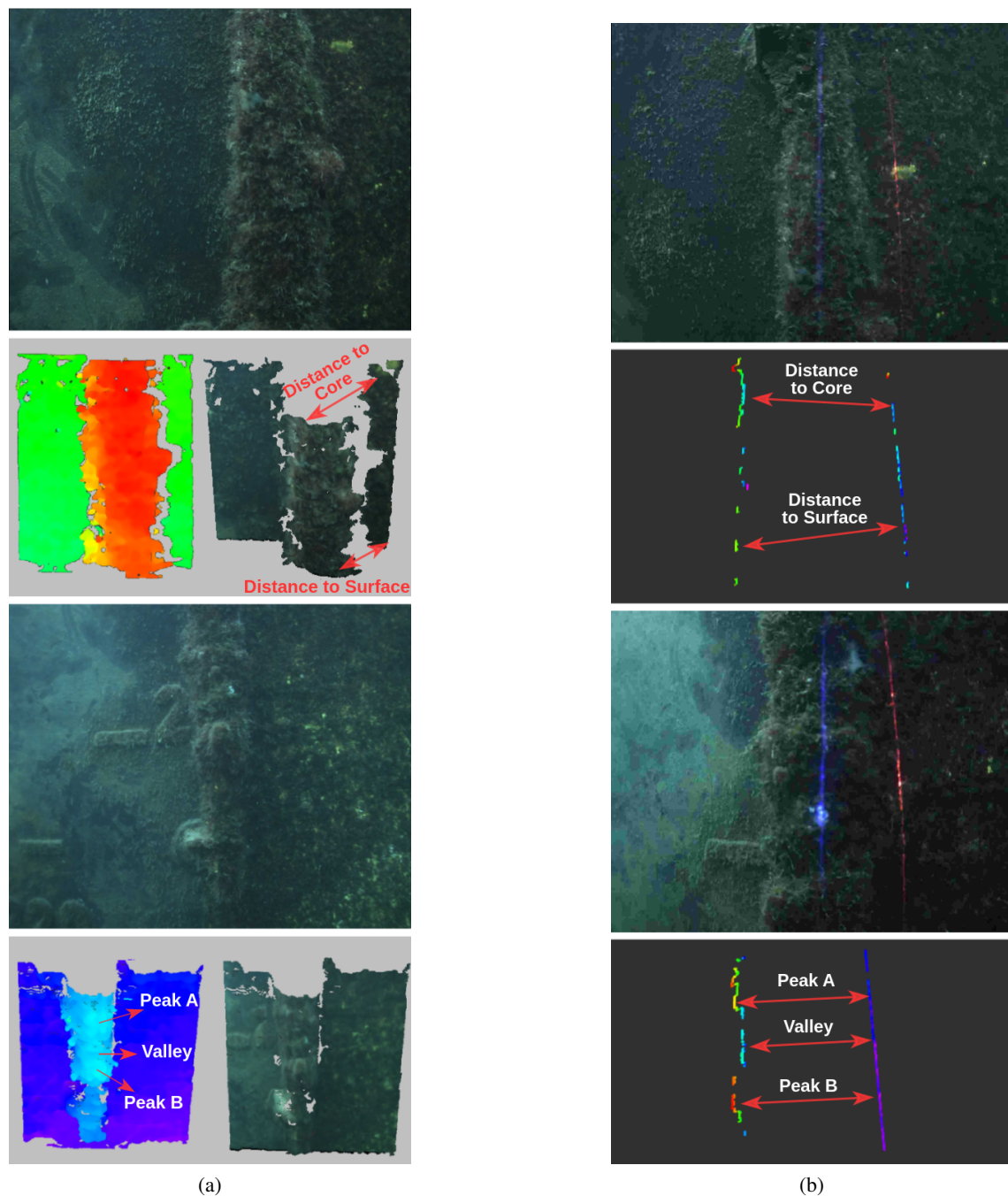


Figure 3.19: MARESyE 3D reconstructions within a real scenario at the ATLANTIS Test Center: photogrammetric stereo (a) and light stripe ranging (b) point clouds and corresponding reference frames.

measurements. The LSR reconstruction error is thus mostly guided by the laser line projected onto the anode's surface. A noticeable discrepancy can be seen in the measurements presented in table 3.9, with the peak characteristics seemingly having much higher absolute errors while the more eroded portions of the anode show millimetric accuracy. These results are however misleading, due to the prior knowledge that the presence of biofouling is not considered in the ground

truth data. This leads to the conclusion that the error inherent to the LSR measurements is being masked out by this extra layer of debris. Referring once more to the experiments presented in table 3.6, a rapid increase in the relative depth measurements' *MAE* can be seen as the sensor gets progressively more distant from the target. Since the tri-dimensional data was collected at circa 0.60 m from the closest anode point (Peak B), these characteristics are estimated with millimetric accuracy and thus most of the absolute error can be considered in fact the thickness of the biofouling in said peaks. As the highly eroded sections of the anode are inherently farther away from the sensor, and the LSR algorithm estimates progressively reduced relative depth measurements due to struggling with smaller laser line discontinuities, both top characteristics along with the anode's bottom valley measurements (table 3.9) have an expected *MAE* of 0.0185 m which is being masked by the aggregation of materials. A rough estimation of this layer of biofouling can thus be estimated by taking the difference between the expected *MAE* and the estimated absolute error for each characteristic. The top section of the anode sees an average debris layer of 0.0143 m, while the amount of biofouling aggregated to the bottom anode valley is assessed to be 0.0086 m thick.

The ATLANTIS Coastal Testbed experiments provide an environment where the sensor receives natural light, as the inspection of the DURIUS structure is conducted near the water surface. However, continuous changes in light conditions result in uneven illumination of the scene, causing contrast limitations, lens flares, and bright spots that hinder the pixel matching process during the PS estimation. The accumulation of biofouling and sediments on the sacrificial anode's surface difficult the analysis of the degradation level of the cathodic protection system, inducing the need for ROVs equipped with cleaning tools (e.g., brushes, water jets) [255]. From an imaging perspective, the textured surface benefits the PS estimation but limits the effectiveness of active techniques, as the laser beams can be easily absorbed. Additionally, moving objects in the scene can cause occlusions that obstruct the acquisition of high-fidelity 3D information. Previous experience allows for a sense of how the performance of the sensor is affected in the low water transparency situations found near the seabed area, where the presence of organic particles, sand and larger living organisms have a large impact in quality of the acquired data. The MARESyE concept has been tested in a controlled scenario to assess its response to photogrammetric phenomena induced by these scenarios [87]. A 90.9 mg L^{-1} concentration of gray mortar dissolved in water emulates high turbidity conditions. MARESyE shows a maximum performance decay of 13 dB in terms of Peak Signal-to-Noise Ratio (PSNR) and 0.28 Structural Similarity Index Measure (SSIM). The hybrid sensor maintains a SSIM index above 0.655, while a conventional imaging system captures degraded data with 0.328 SSIM [87]. In essence, despite being exposed to harsh sub-sea conditions, the hybrid imaging scheme employed by the MARESyE sensor is still capable of capturing reliable information. MARESyE employs a data-driven range-gated system to mitigate the limitations presented by real scenarios, proving its applicability for sacrificial anode inspection procedures where millimetric sub-sea reconstructions are key.

3.6 Final Considerations

The research introduced along this chapter details the design of a hybrid underwater imaging sensor. MARESyE is meticulously thought out to excel in close-range inspection and reconstruction applications. The optical prowess of the system is enabled by mitigation strategies that reduce the impact of sub-sea phenomena. High-fidelity data can thus be gathered from the scenario, avoiding the severe data degradation experienced by off-the-shelf vision-based sensors. MARESyE further contrasts from existing state-of-the-art solutions by retrieving heterogeneous tri-dimensional information through both passive and active imaging techniques. Numerous experiments conducted in a controlled underwater environment validate the initial premise that MARESyE can be employed to capture tri-dimensional information with millimetric accuracy in close proximity to structures of interest - the $[0.5, 0.6]$ m distance spectrum provides the optimal workspace range. The MARESyE sensor is later used in real inspection operations to assess the degradation level of sacrificial anodes. These operations are crucial to protect and extend the lifespan of offshore platforms. In a first stage, with low turbidity and controlled hindrances, the volumetric estimations of a decommissioned target anode are accurate down to an absolute error of 0.0008 m. The final set of experiments are conducted at the ATLANTIS Coastal Test Center, a near-real environment with access to a repurposed floating buoy. MARESyE is integrated into the ROVINGBAT commercial ROV, and guides the inspection mission by providing live visual and metric feedback to the operator. A biofouling covered anode is thoroughly reconstructed and its assessment reveals a layer of approximately 0.0130 m thickness, with peaks and valleys accounting to 0.0143 m and 0.0086 m, respectively. In essence, the MARESyE hybrid imaging system proves capable of enhancing the remote sensing capabilities of AUVs and ROVs by providing textured and precise tri-dimensional information in harsh sub-sea environments, crucial for reliable, repeatable, and dependable O&M procedures.

Chapter 4

Tri-dimensional Information Fusion for Inspecting Submerged Critical Structures

4.1 Introduction

The MARESyE hybrid imaging sensor enables high SNR information to be gathered from the submerged scenario. Both passive photogrammetric stereo and active light stripe ranging techniques are exploited through a range-gated, data-driven strategy to obtain heterogeneous tri-dimensional data streams. Despite mitigating sub-sea degradation effects, the resulting tri-dimensional information can be imperfect and inconsistent since each imaging technique captures data with distinct error profiles, uncertainties and data characteristics. While passive methods retrieve dense and textured information that provide a broad coverage of the scenario, these are often blocky, deeply affected by the lack of features, and its resolution degrades with the distance from the sensor. In turn, LSR beam projections achieve millimeter-precise volumetric measurements through triangulation; however, the sparse nature of this information is insufficient to portray a complete perception of the environment.

This chapter introduces two novel algorithms for fusing heterogeneous 3D data into a unified representation that preserves the advantages of each methodology. The Joint Masked Regression (JMR) approach propagates precise LSR measurements across homogeneous regions surrounding each laser beam projection, using regression curves to refine the stereo-based depth estimates. The learning-based underWater HEterogenous tri-dimensional information fusion Architecture (RHEA) employs early fusion to jointly learn features from coupled 3D inputs, using a synthetic-to-real training scheme that enables direct underwater deployment without domain adaptation. To the best of the author's knowledge, this research presents the first approach to fusing dense, texture-based 3D information with precise but sparse laser-triangulated measurements for underwater structural inspection applications. This chapter is organized as follows: section 4.2 introduces the JMR algorithm, while section 4.3 details the RHEA architecture and its synthetic-to-real

training methodology. An extensive analysis of the performance of both methods is discussed in section 4.4, detailing experiments in simulation, controlled underwater environments and within a real application. Finally, section 4.5 clarifies the most important conclusions of this research, providing a critical view of the obtained results.

4.2 Joint Masked Regression Algorithm

The JMR methodology explores the fundamental principle that homogeneous regions around each segment of the laser projections generally translate the same depth. The very accurate sparse LSR data is used to estimate multiple regression curves that map depth displacements and correlate overlapping information from the PS input to each LSR beam segment. By exploiting the homogeneous regions surrounding each beam, the estimated regression models can be used to correct overlapping pixels and to generalize this information towards neighboring depth values. In summary, the JMR maintains the density of the PS information, while using the triangulated laser data to correct it.

A step-by-step illustration is depicted in figure 4.1, describing each key stage of the JMR algorithm:

1. Find homogeneous regions within the PS 2D projection;
2. Calculate seeds that represent each LSR beam segment projection;
3. Correlate each seed to a homogeneous region within the PS 2D projection through a region growing procedure;
4. Fit a regression curve that maps PS - LSR depth displacements within the homogeneous region surrounding each unique laser beam segment;
5. Correct the PS point cloud by regressing the pixels within each homogeneous mask according to the corresponding polynomial curve.

Initially, both input point clouds must be projected into a two-dimensional image plane. The 2D projection of a tri-dimensional point $P = (P_X, P_Y, P_Z)^T$ can be obtained following the pinhole camera model. Pixel $p = (x, y)^T$ is given by equations 4.1, 4.2 and 4.3 [108]:

$$x = \frac{f_x \cdot P_X}{P_Z} + c_x, \quad (4.1)$$

$$y = \frac{f_y \cdot P_Y}{P_Z} + c_y, \quad (4.2)$$

$$I(x, y) = P_Z, \quad (4.3)$$

where f_x and f_y are the focal length values; c_x and c_y translate the camera's principal point; and I is the intensity at pixel $p = (x, y)^T$, which represents depth in meters. Valid depth readings

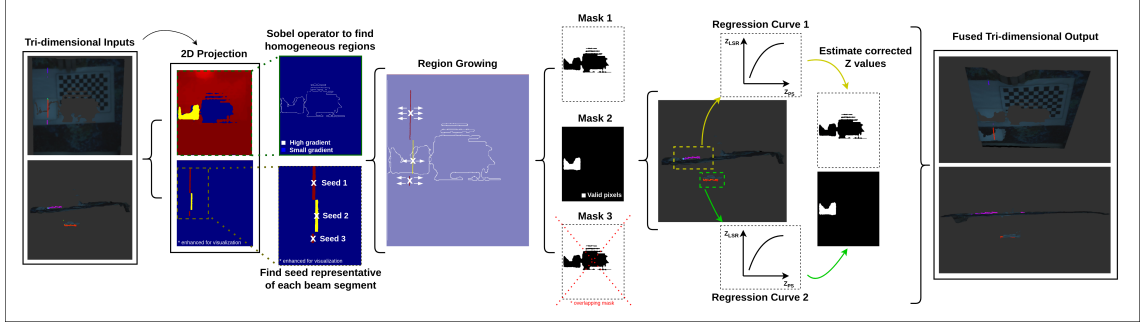


Figure 4.1: Visual representation of the joint masked regression algorithm: sparse laser-based information is expanded towards the homogeneous regions in the PS projection through a region growing procedure; a regression curve is estimated from the overlapping information within each mask and used to correct the stereo-based data.

are considered within the $[1e^{-5}, 2.0]$ m range. Pixels with values outside this range or with no information are assigned intensity 0.

With both point clouds projected into a 2D plane, the homogeneous regions are estimated in the PS projection through a Sobel operator. This 2D kernel operation computes a gradient value for each PS projected pixel. Gradients close to edge and boundary regions are larger than those in homogeneous areas [283]. A pixel value superior to a certain threshold is assigned intensity 255 and considered to belong to a high gradient neighborhood; while the contrary reveals a pixel within a homogeneous region that is represented by intensity 0. This results in a binary gradient mask that delimits homogeneous regions.

Simultaneously, a seed representative of the projection for each beam segment is calculated through a connected components methodology. A binary mask¹ containing the valid readings from the projected LSR input is assigned labels for each set of connected pixels - corresponding to each laser beam segment. The centroid of each individual component is taken as the reference seed for each beam. Each distinct seed is used as starting point for a region growing algorithm that uses the aforementioned binary gradient mask as baseline and expands towards its homogeneous regions. A new mask is generated for each beam segment, translating the areas towards where the LSR information can be expanded - refer to figure 4.1. The high gradient pixels act as a boundary to stop the growing procedure and are not included as part of the generated homogeneous masks.

Each mask M can be represented as a set of m number of pixels p :

$$M = \{p_0, \dots, p_m\}, p_i \neq p_j \forall i, j \in [0, m] \wedge i \neq j. \quad (4.4)$$

The collection of h number of generated masks (H) is given by equation 4.5. Overlapping masks are discarded, assuring that no points are corrected multiple times.

$$H = \{M_0, \dots, M_h\}, M_k \not\subset M_l \forall k, l \in [0, h] \wedge k \neq l. \quad (4.5)$$

¹ A pixel with valid depth reading is assigned an intensity of 255, while a padding pixel is kept with 0 intensity.

The valid depth readings for a single mask M can be formulated for the photogrammetric stereo projection - PS , and light stripe ranging projection - LSR through equations 4.6 and 4.7, respectively.

$$Z_{PS} = PS \cap M = \{z_{PS_0}, \dots, z_{PS_m}\}, z_{PS_t} \in \mathfrak{R}^+ \forall t \in [0, m], \quad (4.6)$$

$$Z_{LSR} = LSR \cap M = \{z_{LSR_0}, \dots, z_{LSR_m}\}, z_{LSR_w} \in \mathfrak{R}^+ \forall w \in [0, m]. \quad (4.7)$$

For the set of homogeneous regions H , the corresponding metric information can then be expressed as:

$$PS_{JMR} = \{Z_{PS_0}, \dots, Z_{PS_h}\}, \quad (4.8)$$

$$LSR_{JMR} = \{Z_{LSR_0}, \dots, Z_{LSR_h}\}. \quad (4.9)$$

Finally, a polynomial regression curve - f_η - is fit to the overlapping depth readings from each of the input point clouds, capturing the depth displacements between the heterogeneous data within each of the masks. Equation 4.10 translates this formulation:

$$f_\eta = \sum_{\epsilon=0}^{\epsilon=p} C_\epsilon PS_{JMR_\eta}^\epsilon, \eta \in [0, h], \quad (4.10)$$

where C_ϵ represents the polynomial coefficients and p is the degree of the curve. This implementation estimates a regression curve with $p = 2$.

The JMR algorithm then uses each curve to correct the PS tri-dimensional points that belong to the corresponding mask, obtaining a single dense and accurate representation of the scenario - refer to equation 4.11. Any points outside the h homogeneous regions are not regressed, remaining intact within the final JMR output - PS_α .

$$JMR_{Fusion} = \{PS_\alpha, f_0(PS_{JMR_0}), \dots, f_h(PS_{JMR_h})\}. \quad (4.11)$$

4.3 Underwater Heterogeneous 3D Information Fusion Architecture

Learning-based methodologies can also be employed in combining heterogeneous tri-dimensional information. The RHEA network makes use of convolutional layers that are responsible for extracting high-level characteristics from the input. The successive use of these layers further compresses these features into a set of low-resolution feature maps. This information is then merged and upsampled until a full-scale resolution depth map is obtained, fusing both the textured stereo-based information with the very accurate triangulated set of points from LSR. In essence, RHEA follows the encoder-decoder architecture pattern represented in figure 4.2.

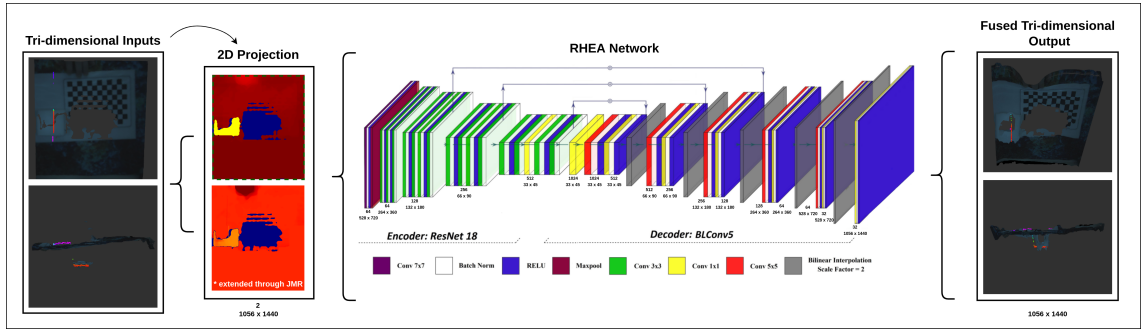


Figure 4.2: Visual representation of the RHEA network: an encoder-decoder architecture that combines both 3D inputs through an early-fusion approach. Feature maps are presented as Channels $\times$ Height $\times$ Width.

As a CNN-based network, RHEA implicitly requires both tri-dimensional inputs to be projected into a 2D plane - refer to equations 4.1, 4.2 and 4.3. After the projection, pixels with valid depth information are asserted into the $[1e^{-5}, 2.0]$ m range. Pixels outside this interval are initialized with value 0 m. Furthermore, the images resulting from the projection are asserted to a size that can be handled by the network (divisible by 16) - 1056×1440 ; the LSR input information is expanded through a slight variation of JMR algorithm where regressed points are added to the LSR point cloud, instead of replacing the corresponding points in the PS input. Equation 4.11 is replaced by equation 4.12, reflecting this variation. The original points within the LSR point cloud are not altered - LSR_α - while the regressed information replaces the previously missing depth readings in the LSR projection.

$$LSR_{extended} = \{LSR_\alpha, f_0(PS_{JMR_0}), \dots, f_h(PS_{JMR_h})\}. \quad (4.12)$$

The fusion procedure is conducted as an early-fusion process, i.e. both inputs are combined *a priori* and features are conjointly learned as the coupled input reaches the encoder - ResNet18 [284]. Its residual connections mitigate the vanishing gradients problematic while allowing for an easier optimization due to the improved flow of gradients. Additionally, the small complexity of the network allows inference to be conducted directly within the MARESyE sensor, enabling edge computing. The first convolutional layer of the ResNet18 is altered to support a 2-channel input containing a coupled representation of the 2D projections of both tri-dimensional input data obtained through the PS and LSR modules. The softmax layers related to the original classification task are removed. A final convolutional layer is introduced to assert the number of channels of the resulting feature map, while maintaining spatial dimensions, so that it can be processed by the decoder - refer to figure 4.2. At this stage, the extracted features are compressed and encoded into low-resolution embeddings. The upsampling operation is conducted through the BLConv5 decoder [42]. Five sequential blocks composed of depthwise convolutions and bilinear interpolations expand the spatial dimensions of the feature map by a factor of 2, while compressing it channelwise. The use of depthwise convolutions contributes towards a lightweight model as these factorize the standard layer into two distinct operations: a first convolution applies the kernel

filter that spatially merges the feature maps; while a second pointwise convolution combines the outputs, resulting in a lower complexity operation [285]. Additive skip-connections are employed between both ends of the architecture to allow high-level features from the encoder to be merged with low-level ones during the upsampling process. This strategy leads to more detailed depth maps, however, it can also convey unwanted information such as error patterns contained in the input data. To avoid this, skip-connections are not implemented in the first convolutional layers.

The lack of disseminated underwater datasets, in particular containing 3D information, presents one of the biggest challenges when it comes to deploying learning-based methods in real sub-sea scenarios [44]. To bypass the need for transfer learning or similar domain-adaptation techniques, this chapter exploits a synthetic-to-real supervised training procedure. The heterogeneous tri-dimensional information gathered by the MARESyE sensor is thoroughly simulated in Gazebo². Training RHEA with data that not only resembles real underwater 3D information, but also accounts for the harsh environmental conditions that degrade the quality of the obtained data, allows it to be directly employed within underwater contexts without the need for re-training. MARESyE's simulated sensor placement and relative orientations follow the original external calibrations. The LSR information is simulated via a 2D laser that mimics the sparse narrow line of 3D points, while its FOV is adjusted to match the practical spatial coverage. Gaussian noise with millimeter levels of uncertainty further approximates the simulated data to the real tri-dimensional information obtained from the active technique. On the other hand, the dense point cloud is captured by a 128-beam LiDAR. The lack of texture in comparison to the real stereo-based information is not relevant for the training procedure since RHEA is directly trained with non-textured 2D projections that solely represent the depth readings from the gathered tri-dimensional information. The PS image texture is only used to colorize the final output prediction within real underwater applications. Originally, the 3D LiDAR point cloud is vastly more accurate and provides more resolution than the real PS stereo obtained in underwater environments. Therefore, the data collected in simulation is tuned to resemble the noise characteristics and depth resolution of the PS point cloud. Furthermore, since the PS module employs an SGBM algorithm [140] to find pixel correspondences, the synthesized stereo-based point cloud aims to mimic some of the common challenges faced by these methodologies: block heavy appearance, sensible to parameter tuning and calibration accuracy, directly affected by dynamic environments that introduce large amounts of floating particles and illumination changes. The synthesized stereo input is thus modeled empirically from characteristics acquired from practical experimentation in diverse scenarios ranging from controlled to very harsh underwater environments:

- The 2D projection of the point cloud is digitized into 20 bins with 0.05 m resolution. The need to compress and resize images to reduce latency in pixel correspondence leads to lower resolution point clouds as well as the blocky appearance common from algorithms employed in stereo-based calculation;

²<https://gazebo-sim.org/home>

- A random offset within the $[-0.05, 0.05]m$ range is added to the depth readings of the point cloud; Poor parameter tuning for the SGBM algorithm, inaccurate stereo calibrations or even small mechanical vibrations during operation can lead to systematic deviations within the PS point cloud;
- Gaussian noise is introduced to each 3D point. Combinations with mean $\mu \in \{-0.25, 0, 0.25\}m$ and standard deviation $\sigma \in \{0, 0.015, 0.03\}$ generate multiple noise masks for each point cloud, augmenting the amount of distinct instances. Floating particles, bubbles and sudden illumination changes typical of underwater scenarios induce constant discrepancies into the pixel matching operation. This results in erroneous information being introduced into the PS point cloud;
- Periphery pixel areas are commonly more affected by radial distortions introduced by the camera lens [286]. This leads to pixel correspondences being more prone to errors when further away from the sensor's center of projection. To mimic this distortion, noise values are scaled through an inverse distance transform that derives a pixel's distance to the center of the image. Each pixel is assigned a value in the $[0, 1]$ range according to its distance to the center of projection. The further away from the center, the higher the assigned value. Noise masks are then multiplied by these distance values to ensure a higher concentration of inaccurate pixels within the periphery of the point cloud.

Some relevant objects from the ModelNet40 dataset [287] are used as targets from which the simulated version of MARESy captures 3D information, amounting to 32000 and 8000 instances for training and testing, respectively. Optimization of the network is carried out through a custom l_2 -based function that can be divided into three components. The global loss function is aggregated through equation 4.13.

$$l_2_{composed} = 0.75(l_2_{valid}) + 0.05(l_2_{invalid}) + 0.20(l_2_{nbsmoothness}). \quad (4.13)$$

Each loss term is expressed in equations 4.14, 4.15 and 4.16:

$$l_2_{valid} = (i - \hat{i})^2, \quad \forall i \in [1e^{-5}, 2.0]m, \quad (4.14)$$

$$l_2_{invalid} = (i - \hat{i})^2, \quad \forall i \in \{0\}m, \quad (4.15)$$

$$l_2_{nbsmoothness} = (AvgPool_{2D}(i) - AvgPool_{2D}(\hat{i}))^2, \forall i \in \{0\} \cup [1e^{-5}, 2.0]m, \quad (4.16)$$

$$AvgPool_{2D}(i(Batch, Channels, Height, Width)) = \frac{1}{k * k} \sum_{g=0}^{k-1} \sum_{r=0}^{k-1} i(Batch, Channels, stride[0] \times Height + g, stride[1] \times Width + r), \quad (4.17)$$

where i and $\hat{i}$ are the ground truth and estimated prediction, respectively. The $AvgPool_{2D}$ function follows equation 4.17 with $K=(3,3)$, $Stride=(1,1)$ and $Pad=0$.

Pixels that represent valid depth information are heavily penalized through term $l2_{valid}$, constituting 75% of the loss function. These pixels correspond to tri-dimensional points that are already contained within the input information, having a direct correspondence to the faithful ground truth data. As the network mainly focuses in fusing this heterogeneous 3D information to effectively output a more precise point cloud, this term mostly drives the optimization process. The $l2_{nbsmoothness}$ term is introduced to promote neighborhood consistency, accounting for 20% of the loss function. The goal is to guide the network to not focus solely on pixel-to-pixel base regression, but also to take into account the surrounding neighborhood. This enforces consistency within the prediction since each pixel is being penalized not only with basis on how well it translates depth, but also how well inserted it is within its corresponding neighborhood of pixels. Finally, the $l2_{invalid}$ term penalizes the network for introducing erroneous information into areas of the 2D plane that only serve as padding to the point cloud projection. A small weight of 5% proves enough to guide the network to learn which areas are purely padding - this is further enforced by the $l2_{nbsmoothness}$ term. Additionally, the low emphasis on this term allows the network flexibility to add new information within regions contained in the neighborhood of pixels with valid depth readings. By having limited impact in the loss function, pixels previously considered invalid can now translate faithful information without the network driving them back to intensity 0. This allows for the densification of the output prediction, possibly filling in previously occluded zones by utilizing context and neighboring information.

4.4 Results and Discussion

The content of this section aims to thoroughly discuss and analyze the performance of the proposed fusion methodologies introduced along this chapter. First, experiments conducted with synthetic data (section 4.4.1) demonstrate that both the joint masked regression algorithm and the RHEA learning-based approaches are capable of handling noisy input data, outputting a vastly improved point cloud both metrically and qualitatively when compared to a given baseline. Nevertheless, underwater data poses extended challenges that are difficult to simulate such as the impact of floating particles, lighting conditions etc. Therefore, section 4.4.2 explores the effectiveness of the proposed synthetic-to-real training method by evaluating the precision of the fused tri-dimensional output within a controlled environment - a clean water tank. Additionally, extensive trials are also

conducted at the ATLANTIS Coastal Test Center [6] to assess both fusion procedures within a real underwater environments - section 4.4.3.

Overall, the quantitative discussions are guided by the following metrics, used to evaluate the output predictions:

- **MAE** - Mean Absolute Error.
- **RMSE** - Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{n} \sum_{k=1}^n (i_k - \hat{i}_k)^2}. \quad (4.18)$$

- δ_n - Percentage of predicted pixels whose relative error is within a threshold - refer to equation (4.19) [121]:

$$\delta_n = \frac{\text{card}(\{\hat{i} : \max\{\frac{\hat{i}}{i}, \frac{i}{\hat{i}}\} < 1.25^n\})}{\text{card}(\{i\})}, \quad (4.19)$$

where i and $\hat{i}$ are the ground truth and prediction, respectively; *card* represents the cardinality of a set.

The models generated by training the RHEA learning-based approach are obtained by using the following hardware: NVIDIA GeForce RTX 2080 SUPER with 8GB of VRAM, Intel i5-8600K CPU @ 3.60GHz with 6 cores. Implementation of the network is conducted in Pytorch. Optimization is carried out through the Stochastic Gradient Descent (SGD) algorithm, with a learning rate of 0.1, decreased by a ratio of 1/10 every 3 epochs. Each training procedure is conducted with batch size 6, during 30 epochs. A weight-decay of $1e^{-3}$ is applied to ensure regularization.

4.4.1 Synthetic Data Experiments

The lack of available underwater datasets composed by heterogeneous tri-dimensional information difficults the use of supervised learning strategies within this context. Synthetic data can be used instead to mitigate this challenge. To train models that are viable and applicable in later stages of testing (i.e. real underwater conditions) the emulated inputs should be as close to the real data as possible, hence the synthetic-to-real approach discussed in section 4.3.

To assess the performance of the proposed methodologies, a baseline d) is established by comparing the synthetic stereo input directly with the ground truth. Table 4.1 presents the metric comparison between distinct fusion methodologies and this baseline. The joint masked regression algorithm - approach a) - provides the smallest improvement, marginally increasing the δ_1 percentage. Metrically, however, the RMSE improves by 15.7%. As a regression-based algorithm, the output point cloud remains unchanged where no LSR information is provided - refer to Prediction a) in figure 4.3. Consequently, the majority of the error is concentrated within the regions

Table 4.1: Performance of the proposed fusion methodologies with synthetic data, compared to the estimated baseline.

Approach	RMSE (m)	MAE (m)	δ_1 (%)
a) Joint masked regression	0.0428	0.0338	44.5
b) RHEA network w/ no preprocessing	0.0294	0.0227	96.7
c) RHEA network w/ LSR extended through JMR	0.0172	0.0119	99.0
d) Baseline	0.0508	0.0476	43.5

Lower is better: RMSE, MAE

Higher is better: δ_1

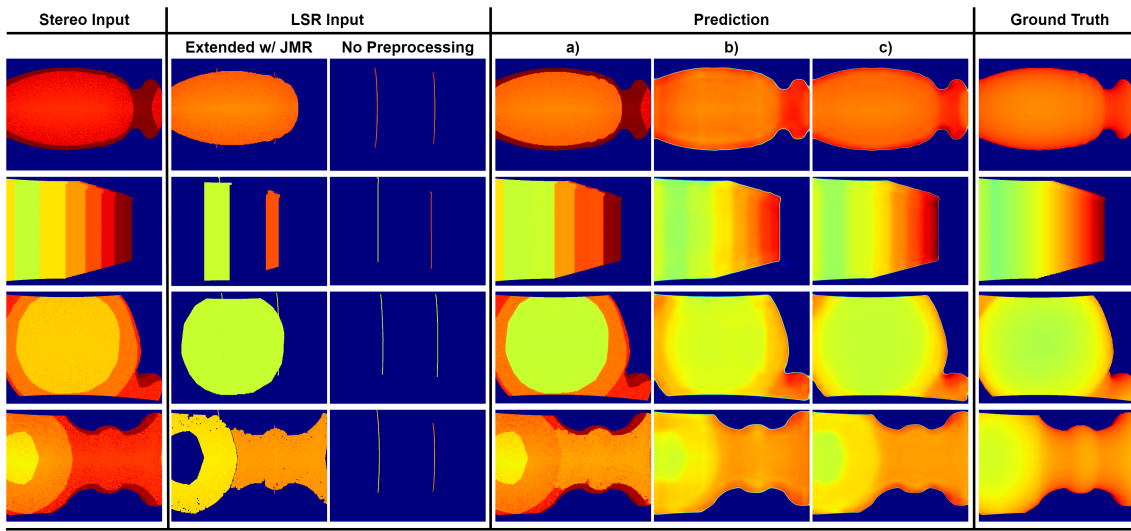


Figure 4.3: Visual comparison between the proposed fusion methodologies using synthetic information. Prediction a) is obtained through the JMR algorithm, using extended LSR information. Prediction b) is inferred by the RHEA network using sparse LSR information with no preprocessing as an input. Prediction c) is estimated by RHEA when using extended LSR information as an input. The stereo input and ground truth are common to every approach.

that lack regression as corroborated by the RMSE map presented in figure 4.4a. The laser stripe positioning is a key factor in the performance of the JMR algorithm, since the output point cloud is directly influenced by how much information can be conveyed by the LSR input.

Learning-based methodologies prove to be more flexible in this aspect, being able to improve the input point cloud even with limited information. Such is the case of approach b) - figure 4.3. The network is capable of translating the features captured from the sparse input to the surrounding neighborhood pixels, achieving a considerable improvement in terms of RMSE - 0.0294 m, and MAE - 0.0227 m; while reaching a δ_1 of 96.7%. On the downside, the neighborhood consistency degrades, leading to artifacts in the prediction and rougher gradient patches. Additionally, the edge regions accumulate the highest amount of error as seen in figure 4.4b. Conclusions drawn from these experiments, align with P. N. Leite *et al.* (2020) [43] in the sense that sparse laser information is only capable of guiding the network to a certain extent, as the stereo input data

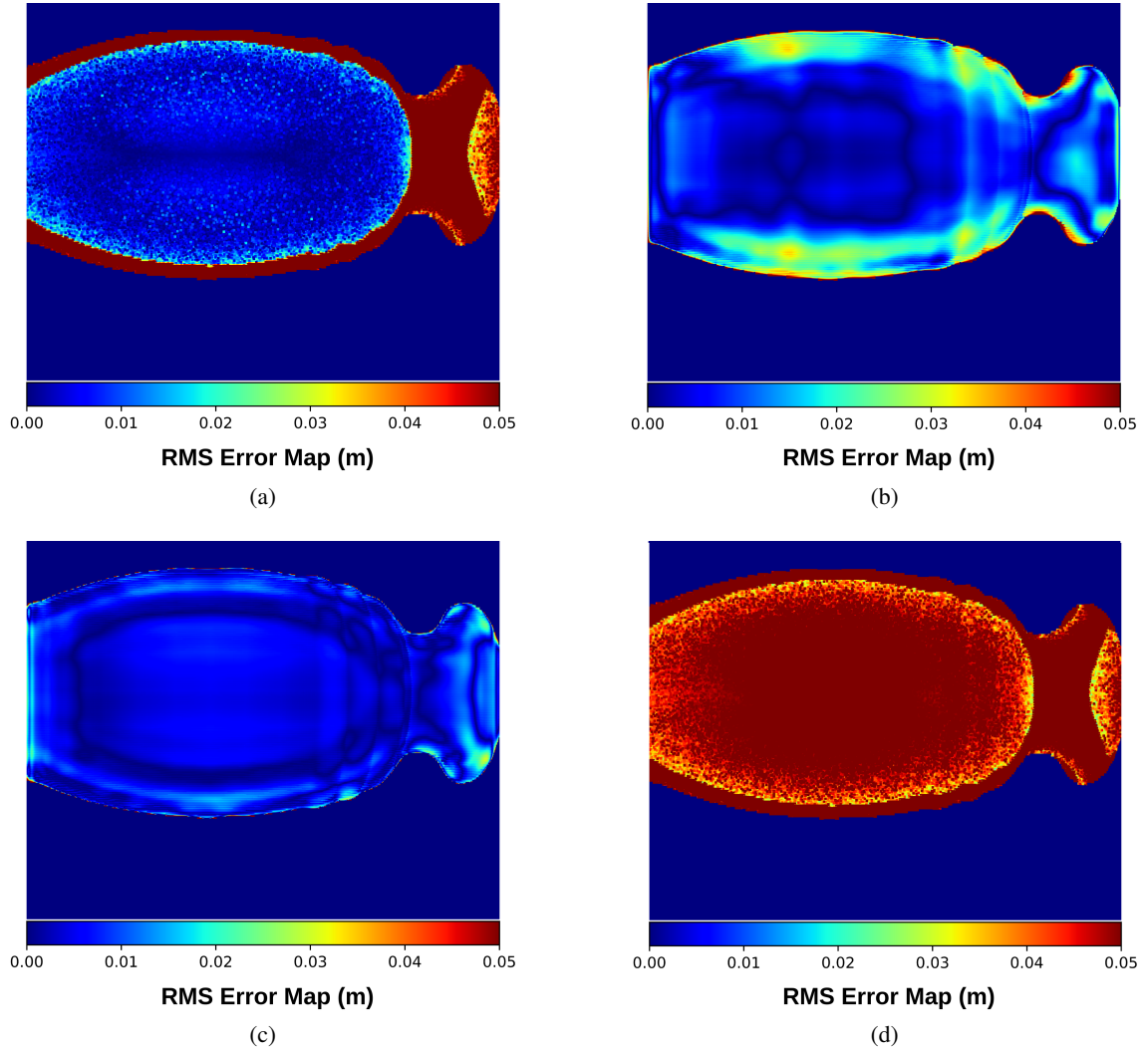


Figure 4.4: RMSE maps obtained by comparing the respective Predictions a), b) and c) - refer to the first row of figure 4.3 - to the ground truth. The baseline error map d) is calculated by taking the difference between the stereo input and the ground truth.

still has major influence in the output resulting in the discussed drawbacks. Aiming to foster point cloud improvement but also preserve its cohesion, approach c) combines the previous two by using the same network architecture as b), however trained with the extended LSR input data obtained through the JMR algorithm. The resulting model vastly improves the stereo input point cloud, surpassing both approach a) and b). The RMSE decreases to 0.0172 m accounting for a metrical improvement of over 66%, while maintaining a very high δ_1 percentage. Prediction c) of figure 4.3 shows much smoother gradient transitions, as well a more consistent pixel neighborhood when compared to Prediction b). Moreover, figure 4.4c reveals an improvement in the pixel predictions near the edges of the point cloud. Expanding the LSR input has a rather positive effect on the network, since it has more faithful information to draw features from; allowing the network to focus on regressing areas that solely rely on stereo input.

Overall every approach is successful, on average, at producing a more precise and faithful output point cloud. The positioning of the laser beams whilst projected on the scene directly impacts the joint masked regression task, since it fundamentally depends on the quality and resolution of the LSR input. On the other hand, deep learning approaches are more flexible and capable of handling sparse inputs, showing vast improvements when the JMR algorithm is used as a preprocessing tool.

4.4.2 Controlled Underwater Experiments

The experiments conducted with synthetic data serve as a great first trial for the proposed fusion algorithms. How well the obtained knowledge translates to sub-sea environments depends on how faithful data representation was during the training phases of the network. This section explores the performance of both the joint masked regression algorithm and the RHEA network (using JMR as preprocessing for the LSR input); also validating the designed synthetic-to-real training methodology through a set of experiments conducted on a controlled environment.

A set of objects is defined and used during trials. The lack of ground-truth data underwater is bypassed by using relative measurements of said objects as the target for a metric analysis - refer to figure 4.5. The first experiment sets the MARESyE system and the object in fixed positions relative to each other within a water tank. For each object of the set, point clouds are gathered through photogrammetric stereo and light stripe ranging. These are analyzed to obtain both spatial and volumetric dimensions, as well as the distance to the camera. This point cloud information is then fused together using both algorithms. Output predictions are studied for the aforementioned characteristics and the results are expressed in table 4.2.

The JMR algorithm is consistently capable of fusing both input point clouds to obtain a more faithful representation of each object. Aligned with conclusions drawn from synthetic data, the correct positioning of the laser beams is a key aspect for the performance of this method; and while the LSR input conveys rich information about the volumetric component of each object, the

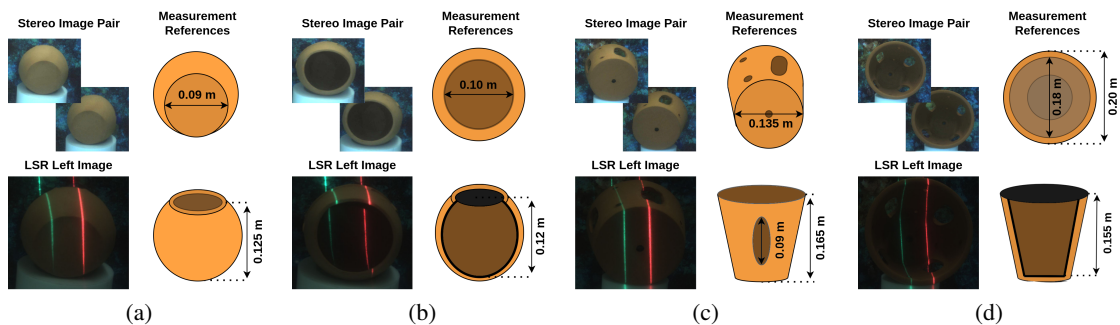


Figure 4.5: Set of objects used during controlled underwater experiments. Characteristics used for analysis are depicted for each object, along with the pair of images gathered during photogrammetric stereo mode (enhanced for visualization) and the corresponding reference image obtained during the light stripe ranging stage.

Table 4.2: Performance comparison between the proposed fusion methodologies using the underwater set of objects depicted in figure 4.5. Instances highlighted in bold represent the highest metric improvement of the input point cloud.

Object	Characteristics		Ground Truth	Photogrammetric Stereo		Light Stripe Ranging		Joint Masked Regression		RHEA Network	
			Measured (m)	Estimated (m)	Absolute Error (m)	Estimated (m)	Absolute Error (m)	Estimated (m)	Absolute Error (m)	Estimated (m)	Absolute Error (m)
A	Spatial Dimensions	Back Diameter	0.0900	0.1030	0.0130	(*)		0.1030	0.0130	0.0995	0.0095
	Volumetric Dimensions	Height	0.1250	0.0370	0.0880	0.1205	0.0005	0.0370	0.0880	0.1284	0.0034
	Distance to Camera		0.4900	0.4968	0.0068	0.4873	0.0027	0.4932	0.0032	0.4980	0.0080
B	Spatial Dimensions	Inner Diameter	0.0100	0.0876	0.0124	(*)		0.0876	0.0124	0.0924	0.0076
	Volumetric Dimensions	Depth	0.1200	0.1442	0.0242	0.1232	0.0032	0.1270	0.0070	0.1317	0.0117
	Distance to Camera		0.5000	0.5049	0.0049	0.5039	0.0039	0.5016	0.0016	0.5044	0.0044
C	Spatial Dimensions	Back Diameter	0.1350	0.1376	0.0026	(*)		0.1376	0.0026	0.1328	0.0022
	Volumetric Dimensions	Height	0.1650		(*)	0.1703	0.0053	(*)		0.1820	0.0170
	Distance to Camera		0.4500	0.4649	0.0149	0.4469	0.0031	0.4419	0.0081	0.4422	0.0078
D	Spatial Dimensions	Inner Diameter	0.2000	0.2109	0.0109	(*)		0.2084	0.0084	0.2174	0.0174
		Outer Diameter	0.1800	0.1863	0.0063			0.1836	0.0036	0.1897	0.0097
	Volumetric Dimensions	Depth	0.1550	0.1929	0.0379	0.1663	0.0113	0.1734	0.0184	0.1748	0.0198
		Ellipse Carving	0.0090	0.1228	0.0328	0.0923	0.0023	0.0867	0.0033		(†)
	Distance to Camera		0.4900	0.5070	0.0170	0.4845	0.0030	0.4930	0.0030	0.4980	0.0080

(*) Laser beam positioning makes it unfeasible to calculate the spatial dimension.
 (*) Not enough information on the point cloud to calculate the volumetric dimension.
 (†) RHEA network introduced new points in the point cloud that completely filled the ellipse carving.

JMR can only make use of this information to correct information that is already contained on the stereo input point cloud, being unable to introduce new points. Such is clear during trials with object A and C, where the laser triangulation estimates the height of the object to be 0.1205 m and 0.1703 m, respectively - table 4.2 - however, the point cloud estimated via photogrammetric stereo is incapable of estimating the full height of the object A (depicted in figure 4.6a), or any point in the area for object C; therefore, the JMR algorithm is unable to use this information - figure 4.6b. When it comes to spatial dimensions, the JMR will consistently mimic the information present on the stereo input, with exception for when the input point cloud lacks consistency around the edge regions. This leads to pixels that should belong to the same surface being included as part of different masks, generated from distinct beam components, and regressed accordingly; resulting in slight deviations to the spatial dimensions - e.g. object D, table 4.2.

The ability to add new information to the input point cloud can be seen as a strong suit of the learning-based approach. Predictions inferred by the RHEA network also represent a metric improvement when compared to the photogrammetric stereo estimation. Being able to use of all the information conveyed by the LSR input, the model is capable of completing portions of the stereo input accordingly. Studying once more the trials with object A, the output prediction fills the missing information allowing a measurement of 0.1284 m for the height of the object, which was previously represented as 0.0370 m in the original point cloud - figure 4.6c. This flexibility of the convolutional neural network, however, can also be detrimental in some cases as it can introduce some unreliable information. Object D - table 4.2 - can be referred to as an example of this downside: the ellipse carving measured as 0.1228 m in the stereo input, is now completely filled in as the network interprets the carving as lack of information, and not as a feature of object D. Moreover, and despite the training procedure enforcing neighborhood consistency, artifacts can still be seen in the predictions, mostly around high gradient regions such as the edges of the point cloud, reflecting the same drawbacks seen in the synthetic data.

These experiments prove the effectiveness of both fusion methodologies in underwater environments. The joint masked regression algorithm is very dependent on stereo input, not only when

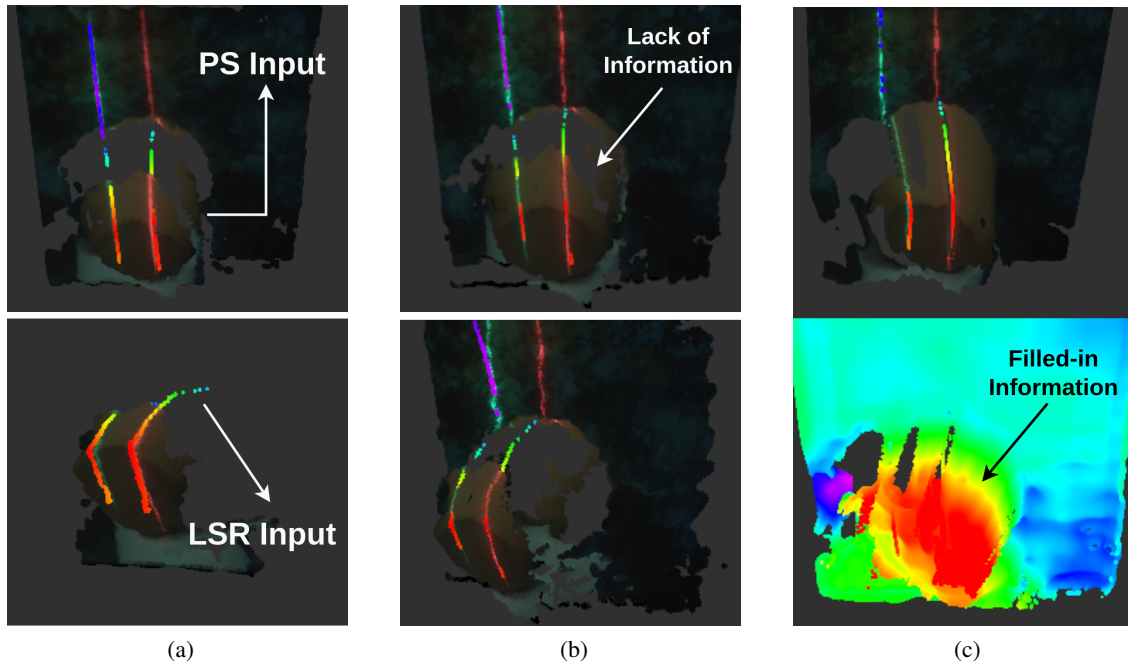


Figure 4.6: Comparison between the distinct point cloud predictions obtained during trials with object A. Figure a) depicts the 3D input information. Figure b) represents the output of the JMR algorithm. Figure c) illustrates the prediction inferred by the RHEA network. Point cloud information is colorized with an image gathered during the LSR stage to prove the tri-dimensional alignment of the heterogeneous input data. The JMR algorithm is unable to add new information to the point cloud, failing to depict the height of the object. RHEA can fill in missing information, fully making use of the LSR input to translate the characteristics of the object.

it comes to spatial dimensions; but also the lack of information derived from occlusion zones. A more precise and faithful output prediction can be obtained when both the stereo and LSR inputs contain overlapping information, struggling otherwise since no new information can be added to the point cloud. The knowledge obtained with synthetic data is correctly conveyed towards its real context application, validating the proposed synthetic-to-real training methodology. The RHEA network is also capable of estimating a metrically improved output point cloud, while able to fill in missing information guided by the LSR input. While resulting in improvements in some cases, these added points are still very susceptible to introducing erroneous information.

4.4.3 ATLANTIS Coastal Testbed - Real Application Experiments

Real underwater environments introduce even more challenges due to the constant changes in lighting, sediments contained in the water that not only hinder visibility but also increase the backscattering of photons, etc. To test the robustness of the proposed fusion methodologies in these conditions, experiments were conducted at the ATLANTIS Coastal Test Center, quayside in Viana do Castelo's Port infrastructure [6]. DURIUS is an offshore floating structure repurposed from the Oil & Gas industry - refer to chapter 3, figure 3.17a. Access to this platform provides the

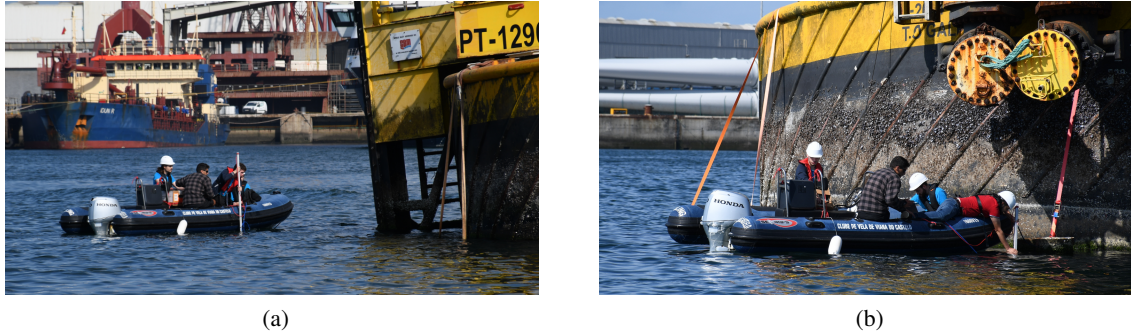


Figure 4.7: INESC TEC team gathering underwater information using MARESyE suspended from a support vessel at the ATLANTIS Coastal Testbed in Viana do Castelo.

perfect opportunity to test MARESyE, as well as the data fusion algorithms in a real application. The sensor was used to gather information in specific areas of interest of the buoy - figures 4.7a and 4.7b. Multiple data collection campaigns were conducted with diverse environmental conditions, causing distinct amounts of suspensoids to be contained within the water. The retrieved heterogeneous tri-dimensional information is later fused together through both the joint masked regression algorithm, and the RHEA network - these results are discussed along this section.

These harsh conditions directly impact the quality of the information gathered by MARESyE: floating particles hinder pixel correspondence for photogrammetric stereo computation; constant shift in light conditions introduces uneven lighting of the scene; absorption of the projected laser beams hampers the triangulation process; etc. The trials conducted with object D, figure 4.5d, were replicated in this environment. Despite the additional challenges of working with less accurate input information, both fusion algorithms are still capable of computing a more precise and faithful dense point cloud. The resulting analysis is presented in table 4.3. Both the joint masked regression method and the RHEA network, show more difficulty in handling the spatial dimen-

Table 4.3: Performance comparison between the proposed fusion methodologies in a real underwater environment, using object D - figure 4.5d. Instances highlighted in bold represent the highest metric improvement of the input point cloud.

		Spatial Dimensions		Volumetric Dimensions		Distance to Camera
		Inner Diameter	Outer Diameter	Depth	Ellipse Carving	
Ground Truth	Measured (m)	0.1800	0.2000	0.1550	0.0900	0.4000
PS	Estimated (m)	0.1743	0.1941	0.1449	0.1010	0.4053
	Absolute Error (m)	0.0057	0.0059	0.0101	0.0110	0.0053
LSR	Estimated (m)	Unable to		0.1602	0.0767	0.3980
	Absolute Error (m)	Estimate (*)		0.0052	0.0133	0.0020
JMR	Estimated (m)	0.1669	0.1868	0.1583	0.1141	0.3982
	Absolute Error (m)	0.0131	0.0132	0.0033	0.0241	0.0018
RHEA	Estimated (m)	0.1692	0.1866	0.1496	Unable to	0.4035
	Absolute Error (m)	0.0108	0.0134	0.0054	Estimate (†)	0.0035

(*) Laser beam positioning makes it unfeasible to calculate the spatial dimension.

(†) RHEA network introduced new points in the point cloud that completely filled the ellipse carving.

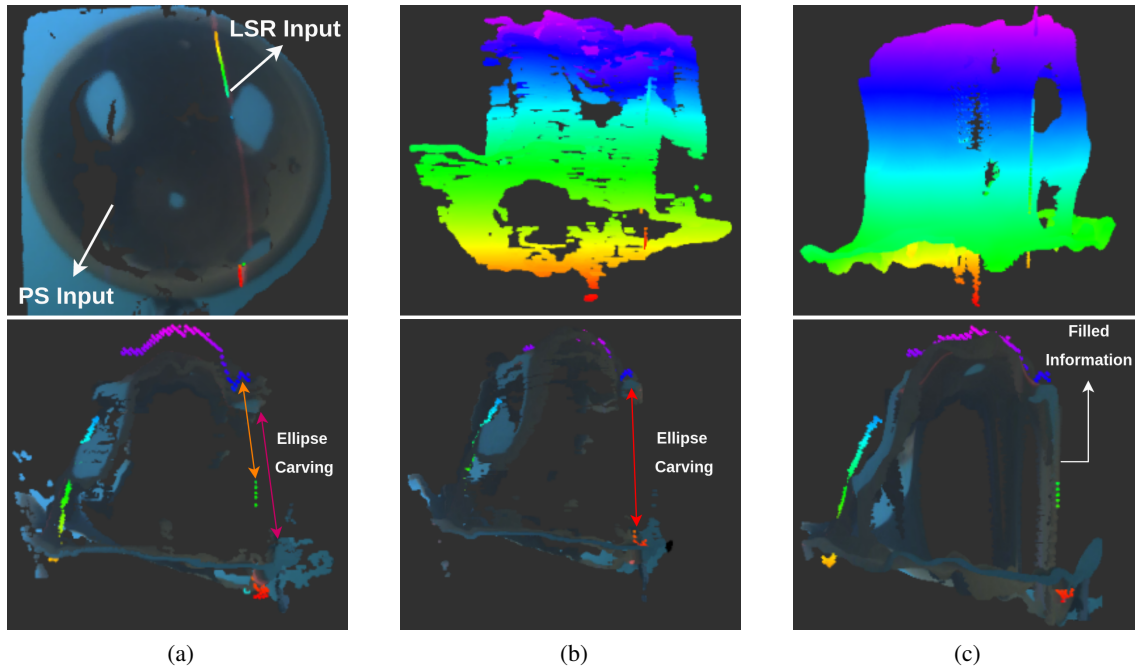


Figure 4.8: Comparison between the distinct point cloud predictions obtained during trials with object D, at the ATLANTIS Coastal Testbed. Figure a) depicts the 3D input information. Figure b) represents the output of the JMR algorithm. Figure c) illustrates the prediction inferred by the RHEA network. Both tri-dimensional fusion algorithms prove capable of handling degraded input data, collected within real underwater environments. The LSR information is used to correct both the global positioning and the volumetric depth dimension of the stereo input.

sions of the object. With no guidance from the LSR input in this regard, and a stereo input that is vastly more prone to having outliers, this outcome can be explained due to faulty mask generation. When expanding the LSR information towards homogeneous regions, these outlier pixels are left out of the generated mask, therefore suffering no alterations; while pixels contained within the mask are corrected according to the estimated regression curve. This leads to small deviations from the measured inner and outer diameters of object D, increasing the absolute error for the spatial dimensions in both predicted point clouds, when compared to the input PS information - refer to table 4.3.

When it comes to the volumetric depth dimension of object D, as well as its global positioning relative to the MARESy camera, both methods prove capable of correcting the input point cloud. The JMR algorithm shows almost perfect alignment with the laser input information - refer to figure 4.8b - resulting in a 0.0033 m error in its representation of the depth of the object. The RHEA model performs slightly worse, nevertheless estimating a more precise output point cloud. The measurements presented in table 4.3 reveal a 0.0054 m and 0.0035 m absolute error for the object's volumetric dimension and distance to camera, respectively. The ellipse carving in object D is a complicated feature to estimate, since it is partly occluded. The information present in the stereo input cloud allows for a length estimation of 0.1010 m - represented by the pink arrow

in figure 4.8a; while the laser information translates a length of 0.0767 m - orange arrow, figure 4.8a. Similarly to the results observed in controlled environments (section 4.4.2), the learning-based solution processes the ellipse carving of object D as lack of information, introducing new points in the area - refer to figure 4.8c. On the other hand, and contrarily to previous results, the JMR algorithm does not improve the input cloud in this dimension - red arrow in figure 4.8b. The carving area does not show a high enough gradient to be considered as part of a different homogeneous region, therefore being included in the same mask as the bottom surface of the object D. Consequently, it is regressed through the same curve, widening the gap as the depth dimension of the object is corrected from 0.1449 m (input point cloud) to 0.1583 m (output of JMR) - refer to table 4.3. In the experiments presented in section 4.4.2 this gap is reduced, since the correction occurs in the opposite direction: the volumetric dimension is adjusted from 0.1929 m (input point cloud) to 0.1734 m (output of JMR) - table 4.2.

Experiments conducted with information retrieved from the DURIUS floating buoy allow the validation of both tri-dimensional fusion algorithms in a real underwater application. Figure 4.9 depicts a few instances from the conducted trials. The lack of ground-truth data limits the discussion to a more qualitative analysis. At the time of inspection, DURIUS was covered with biofouling. These conditions both favor and hinder the tri-dimensional information gathered by MARESyE: while the photogrammetric stereo gains from having such rich texture in the scenario; the LSR is severely affected not only by the absorption of the beam in this sediment heavy water, but also due to the fact that most of the bio organisms have the same coloring as one of the projected laser beams (green - 520 nm), making the segmentation impossible for Instances 1, 2 and 3. Despite having a single laser beam as input, both the joint masked regression and the RHEA network are still capable of using the available sparse information to estimate a more faithful point cloud. Instance 1, figure 4.9 shows the top part of a pillar structure that contains a thick border of mussels attached to it. This is a relatively smooth region upon which the laser information can be expanded. The JMR algorithm makes use of this information to perfectly reshape the stereo input point cloud, mainly the top border of the structure. However, it struggles with the edges of the pillar where it starts to curve into a cylinder-type shape. This is due to the fact that the regression curve is calculated with information from a single beam, positioned at the center of the structure: where there is no curvature; therefore it propagates this information to the remaining points inside the generated mask, ending up overcorrecting the points that start to curve inwards, creating a few outlier points - highlighted in red, Instance 1 - figure 4.9. A continuation of this pillar structure is depicted in Instance 2 of figure 4.9. Once more, the output point cloud estimated through the JMR is corrected in the central portion of the cylinder-shaped structure, but it gets deformed in the edge regions similarly to the results discussed for Instance 1. This can be seen as a systematic downside of the JMR algorithm when there is not enough information to create a faithful regression curve that correctly describes the shape of the object. Additionally, only portion of the stereo input point cloud is regressed, since the outer edges of the pillar are no longer considered as part of the homogeneous region to which the laser information is expanded - highlighted in red, Instance 2 - figure 4.9. The RHEA network shows a higher robustness to working with less information since

it was introduced to scenarios where only information from a single beam was available during its training procedure. This can be observed in its output predictions depicted in figure 4.9. RHEA proves capable of regressing the input point cloud in its entirety despite the very sparse laser information. While both Instances 1 and 2 show an output prediction with corrected global positioning relative to the MARESyE camera, the prediction for Instance 1 introduces a curvature to the point cloud that correctly depicts top border of the cylinder structure, however wrongly propagating this curvature to the remaining portion of the point cloud - highlighted in green, Instance 1 - figure 4.9. Instance 3, figure 4.9 presents a scenario where one of the beams supporting the stairs of DURIUS is overrun with fouling, adding to the already rich depth gradient scenario. In this situation the laser beam information is positioned perfectly to translate part of the volumetric dimension of the structure. Therefore, both tri-dimensional fusion algorithms thrive in this instance, being capable of regressing the input point cloud according to the sparse information. Moreover, the RHEA network introduces points into the output prediction, filling in portions of the point cloud that previously were empty due to being part of occlusion zones, as seen in figure 4.9.

The first data collection campaign with MARESyE in the ATLANTIS Coastal Testbed was successful in retrieving quality information that allowed for the validation of both the joint masked regression and RHEA network tri-dimensional fusion approaches in a real underwater environment and application. However, having a single functional laser had direct impact in the obtained results, mainly affecting the JMR algorithm. With the knowledge from these first trials, the MARESyE imaging system was updated and the green colored laser beam was swapped for a violet colored one (405 nm); to ease laser segmentation and avoid mixing it amongst the existing bio-organisms. The latest campaign occurred under heavy rain and harsh environmental conditions, completely changing the dynamics of the environment since the water becomes vastly more turbid and rich in floating particles. Instance number 4 in figure 4.9 depicts one of the experiments in these conditions, representing a similar cylinder-shaped structure from the DURIUS buoy. In this trial, the JMR algorithm is capable of correcting the input point cloud while maintaining its curved shape, no longer introducing any outlier points in the edge regions. Having two available laser beams leads to this improved outcome, as the regression curve does not have to generalize a single beam to the whole homogeneous region - as it occurs for Instances 1 and 2. The learning-based solution still performs as expected, inferring a point cloud that is both aligned with the LSR input information and also denser as the network introduces new points - refer to the bottom image of Instance 4, figure 4.9. While the points added on the right make sense and help complete the shape of the represented structure, the same cannot be said of the group of points highlighted in green, where the network mimics this behavior and connects both portions of the cylinder incorrectly. These extensive experiments allow for a deep analysis and discussion of the performance of both tri-dimensional fusion methodologies. While the learning-based RHEA network vastly outperforms the JMR algorithm during the trials with synthetic data; the performance of the latter proves very reliable in underwater environments, when the laser beam positioning on the scenario provides enough resolution and correctly translates its volumetric dimension. RHEA is more robust to a sparser set of input features, however being more prone to introducing artifacts in the prediction.

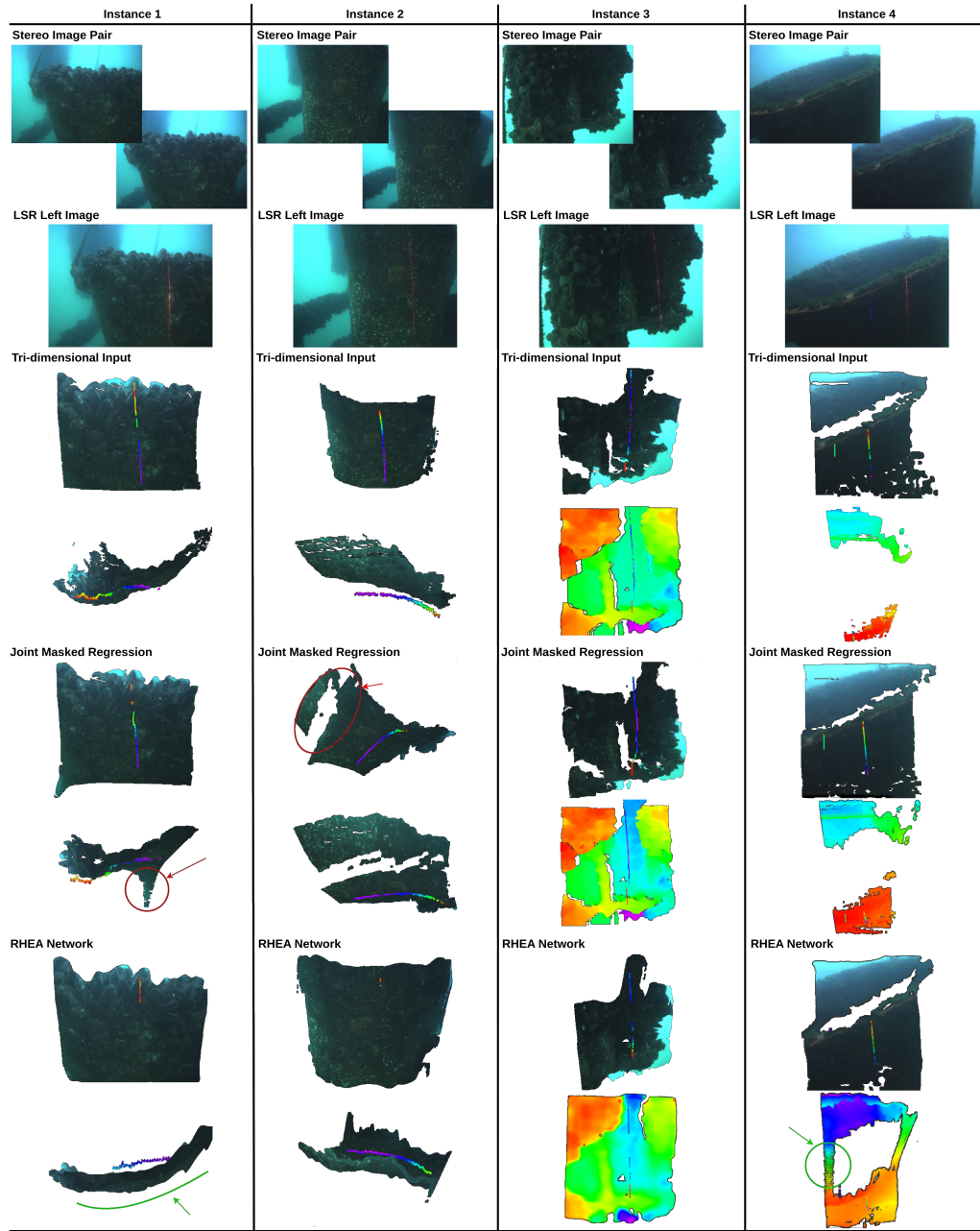


Figure 4.9: Visual comparison between the proposed fusion methodologies using real underwater data collected by the MARESyE sensor at the ATLANTIS Coastal Testbed. Instances 1, 2 and 3 - the acquisition of LSR data is hindered since DURIUS is covered with biofouling, and the green wavelength laser beam blends in with the bio organisms. With a single laser beam as sparse input, the JMR algorithm corrects a portion of the dense point cloud, while introducing some error in the edge regions that depict some curvature (highlighted in red). The RHEA network proves more robust in handling a sparser set of LSR input points. Instance 4 - MARESyE employs a violet colored laser beam during a second set of trials. The additional information improves the performance of both methodologies. JMR correctly estimates the curvature of the structure. RHEA regresses the input data while introducing new points that both fill in occluded zones and introduce outliers (highlighted in green).

Overall, both methodologies are capable of estimating a more precise and faithful output point cloud for this real application, where the harsh underwater conditions degrade the quality of the gathered input data.

4.5 Final Considerations

The heterogeneous 3D information captured by the MARESyE sensor have inherent error distributions and data characteristics: PS provides dense and textured coverage of the scenario with limited precision, while LSR offers millimeter-precise measurements but sparse spatial information. The proposed JMR and RHEA network are the first algorithms capable of fusing heterogeneous 3D data with the aforementioned characteristics. JMR exploits homogeneous regions to map PS - LSR depth displacements through diverse regression curves. These are later used to correct the dense stereo-based information. RHEA is a learning-based method that relies on early-fusion to jointly learn features from a coupled representation of both tri-dimensional inputs. The optimization of the network employs a synthetic-to-real training scheme, bypassing the need for domain-adaptation methodologies and enabling direct deployment in underwater scenarios.

Extensive experiments validate the proposed fusion algorithms in simulation, controlled environments and, in a real application with data collected from DURIUS. The conducted underwater experiments reveal an average absolute error of 0.0097 m and 0.0133 m for RHEA and JMR, respectively. These results correspond to a 52.5% and 34.9% metric improvement when compared to the input PS point cloud. The harsh underwater conditions posed by the experiments conducted at the ATLANTIS Coastal Testbed and the presence of biofouling attached to the DURIUS structure hinder the data acquisition process. RHEA proves capable of predicting an improved output point cloud, even with the sparser set of input features, further validating the employed synthetic-to-real training scheme. The performance of the JMR algorithm proves reliable when the laser beam projection on the scenario conveys enough information to represent target's characteristics. Otherwise, the algorithm estimates a regression curve that misrepresents the complete scenario, leading to a portion of the point cloud being corrected while the remaining points are either deformed or not modified.

These fusion methodologies represent a significant advance in underwater perception capabilities, enabling more reliable and accurate 3D reconstruction crucial for autonomous inspection tasks. By effectively combining the complementary strengths of heterogeneous 3D data streams, these algorithms enhance the spatial understanding of underwater scenes while maintaining millimeter-level precision.

Chapter 5

Fusing Texture, Motion Perception and Sparse Depth for 3D Scene Reconstruction

5.1 Introduction

MARESyE enables enhanced tri-dimensional perception of underwater environments through multiple sensing modalities. While the system’s passive stereo imaging provides dense scene coverage, its performance is subject to typical photogrammetric limitations (detailed in chapter 3). This chapter introduces the fusiNg tEXture with apparent motion information for underwater Scene recOnstruction (NESO) architecture, proposing a novel approach to underwater scene understanding. NESO leverages complementary information sources: rich contextual features from image data [288] combined with relative depth cues from motion analysis, where pixel displacement magnitude correlates inversely with object distance [42]. The proposed encoder-decoder architecture employs attention-based pyramidal skip connections to propagate relevant features through multiple decoding stages. This design enables dense depth predictions that preserve key scene elements while operating directly on raw data, minimizing the need for extensive parameter tuning.

The initial learning-based representation is further refined through fusion with millimeter-precise measurements from active laser imaging. The joint masked regression method is selected for this heterogeneous 3D fusion task due to its ability to generalize across any dense tri-dimensional input, whereas the RHEA network, being trained on synthetic data tuned for stereo characteristics, is less suitable for processing NESO’s output predictions. This multi-stage approach combines the dense and textured spatial coverage of learning-based predictions with the precision of active sensing, enabling robust underwater scene reconstruction even in challenging conditions. This chapter is organized as follows: section 5.2 presents the NESO architecture and its underlying multimodal fusion principles based on motion perception concepts. Section 5.3 provides a comprehensive evaluation of NESO’s performance, including comparisons with state of the art methods and analysis of its generalization capabilities. Moreover, the gains of fusing the

dense NESO prediction with sparse depth are validated within a controlled environment. Section 5.4 summarizes the key findings and their implications.

5.2 Fusing Texture and Apparent Motion Components for Depth Estimation

The pixel displacements between two distinct viewpoints captured from the same static scene translate relevant tri-dimensional information. A relative sense of depth can be extracted from the motion vectors that express these displacements. The motion profile of an observer moving in this static scenario is represented by both translational ($\dot{\Gamma}$) and rotational (Ω) velocities:

$$\dot{\Gamma} = (t_X, t_Y, t_Z)^T, \quad (5.1)$$

$$\Omega = (w_X, w_Y, w_Z)^T. \quad (5.2)$$

A point $P = (P_X, P_Y, P_Z)^T$ represented in the imaging sensor's coordinate space will change its position according to the movement of the observer. The velocity of P , considering that it moves away from the sensor, is thus given by equation 5.3 [289]:

$$\dot{P} = -\dot{\Gamma} - \Omega \times P. \quad (5.3)$$

The two-dimensional velocity vector γ of the projected point can be obtained by taking the derivative of $p = (x, y)^T$:

$$\gamma \equiv \begin{bmatrix} \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial t} \end{bmatrix}. \quad (5.4)$$

Substituting equations 4.1 and 4.2 into equation 5.4:

$$\gamma = \frac{1}{P_Z} \begin{bmatrix} f_x \dot{P}_X - f_x \frac{P_X \dot{P}_Z}{P_Z} \\ f_y \dot{P}_Y - f_y \frac{P_Y \dot{P}_Z}{P_Z} \end{bmatrix} = \frac{1}{P_Z} \begin{bmatrix} f_x \dot{P}_X - x \dot{P}_Z \\ f_y \dot{P}_Y - y \dot{P}_Z \end{bmatrix} = \frac{1}{P_Z} \begin{bmatrix} f_x & 0 & -x \\ 0 & f_y & -y \end{bmatrix} \begin{bmatrix} \dot{P}_X \\ \dot{P}_Y \\ \dot{P}_Z \end{bmatrix}. \quad (5.5)$$

Solving equation 5.5 in order to $\dot{P}$ and substituting it into equation 5.3:

$$\gamma = \frac{1}{P_Z} \begin{bmatrix} -f_x & 0 & x \\ 0 & -f_y & y \end{bmatrix} \dot{\Gamma} + \frac{1}{P_Z} \begin{bmatrix} -f_x & 0 & x \\ 0 & -f_y & y \end{bmatrix} (\Omega \times P). \quad (5.6)$$

Expanding the cross product between vectors Ω and P :

$$(\Omega \times P) = \begin{bmatrix} \Omega_X \\ \Omega_Y \\ \Omega_Z \end{bmatrix} \times \begin{bmatrix} \frac{xP_Z}{f_x} \\ \frac{yP_Y}{f_y} \\ P_Z \end{bmatrix} = \begin{bmatrix} \Omega_Y P_Z - \Omega_Z \frac{yP_Z}{f_y} \\ \Omega_Z \frac{xP_Z}{f_x} - \Omega_X P_Z \\ \Omega_X \frac{yP_Z}{f_y} - \Omega_Y \frac{xP_Z}{f_x} \end{bmatrix} = P_Z \begin{bmatrix} 0 & 1 & -\frac{y}{f_y} \\ -1 & 0 & \frac{x}{f_x} \\ \frac{y}{f_y} & -\frac{x}{f_x} & 0 \end{bmatrix} \begin{bmatrix} \Omega_X \\ \Omega_Y \\ \Omega_Z \end{bmatrix}. \quad (5.7)$$

Vector γ can now be expressed using equation 5.8:

$$\gamma = \frac{1}{P_Z} \begin{bmatrix} -f_x & 0 & x \\ 0 & -f_y & y \end{bmatrix} \dot{\Gamma} + \begin{bmatrix} -f_x & 0 & x \\ 0 & -f_y & y \end{bmatrix} \begin{bmatrix} 0 & 1 & -\frac{y}{f_y} \\ -1 & 0 & \frac{x}{f_x} \\ \frac{y}{f_y} & -\frac{x}{f_x} & 0 \end{bmatrix} \Omega. \quad (5.8)$$

Finally, the two-dimensional velocity of p can be formulated as:

$$\gamma = \frac{1}{P_Z} \begin{bmatrix} -f_x & 0 & x \\ 0 & -f_y & y \end{bmatrix} \dot{\Gamma} + \begin{bmatrix} \frac{xy}{f_y} & -f_x - \frac{x^2}{f_x} & \frac{f_x y}{f_y} \\ f_y + \frac{y^2}{f_y} & -\frac{xy}{f_x} & \frac{f_y x}{f_x} \end{bmatrix} \Omega. \quad (5.9)$$

Equation 5.9 reveals that the apparent motion of the system does not directly depend on the tri-dimensional coordinates of point P , as vector γ is obtained from the sum of the observer's translational and rotational velocity components [289]. Considering an horizontally aligned stereoscopic optical system, the rotational component Ω is null, as both cameras maintain parallel optical axes. Therefore, the instantaneous pixel velocity becomes solely dependent on the translational component:

$$\gamma = \frac{1}{P_Z} \begin{bmatrix} -f_x & 0 & x \\ 0 & -f_y & y \end{bmatrix} \dot{\Gamma}. \quad (5.10)$$

The apparent movement of p can be obtained by integrating the velocity equation:

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \int \gamma dt = \frac{1}{P_Z} \begin{bmatrix} -f_x & 0 & x \\ 0 & -f_y & y \end{bmatrix} \int \dot{\Gamma} dt. \quad (5.11)$$

For a synchronized stereo system, the capture of both camera views occurs at the same instant. The translation T from the left to the right image is instantaneous ($\Delta t \rightarrow 0$), where the velocity becomes a spatial displacement. By applying the Mean Value Theorem [290], the integral over an infinitesimal time interval represents the instantaneous spatial displacement between cameras. Given the horizontal alignment with baseline B :

$$\int_{t_0}^{t_0 + \Delta t} \dot{\Gamma} dt = \gamma(c) \Delta t \xrightarrow{\Delta t \rightarrow 0} T = \begin{bmatrix} B \\ 0 \\ 0 \end{bmatrix}, \quad c \in [t_0, t_0 + \Delta t]. \quad (5.12)$$

The pixel displacements can thus be expressed through equations 5.13 and 5.14:

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \frac{1}{P_Z} \begin{bmatrix} -f_x & 0 & x \\ 0 & -f_y & y \end{bmatrix} \begin{bmatrix} B \\ 0 \\ 0 \end{bmatrix}, \quad (5.13)$$

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} \frac{-f_x B}{P_Z} \\ 0 \end{bmatrix}. \quad (5.14)$$

The scene depth P_Z appears coupled to the horizontal pixel displacements (Δx) as an inversely proportional factor, i.e. an object closer to the observer shows larger apparent movement than an object further away. The optical properties of the imaging system, focal length (f_x) and baseline (B) act as a scaling factor. The zero vertical disparity ($\Delta y = 0$) demonstrates that corresponding points satisfy the epipolar geometry constraint. This geometric relationship can be expressed more explicitly as:

$$\Delta x = \frac{-f_x B}{P_Z}, \quad (5.15)$$

where the inverse proportionality between pixel displacements (disparity) and depth becomes evident.

Figure 5.1 illustrates this through a stereo-based system where both the left and right camera images represent two viewpoints captured in the same instant, ensuring a static scenario. The fish in the foreground (highlighted in red) are close to the camera, as such there's significant movement in between images, hence the corresponding high magnitude of pixel movement. Contrarily, the background wall (highlighted in blue) is considerably further away from the observer, inducing smaller pixel displacements.

A vector field that represents the magnitude and direction of each pixel's movement can be obtained by employing optical flow techniques [291]. Traditionally, the estimation of these motion components assumes that temporal changes to a pixel's intensity are caused by motion: a pixel with intensity $I(x, y, t)$, displaced by δ_x and δ_y during the time interval δ_t , will then be located at $I(x + \delta_x, y + \delta_y, t + \delta_t)$. Optical flow algorithms are hindered by fast-moving objects, occlusions, motion blur, and non-textured surfaces [292]. This research diverts from traditional time-sequential optical flow estimation by analyzing motion vectors between stereo camera viewpoints rather than temporal frames. This approach eliminates temporal dependencies and their associated challenges, focusing instead on spatial displacement in a static environment where pixel movement depends solely on the fixed baseline translation between cameras. The calibrated stereo configuration, with parallel optical axes, ensures null rotational components, allowing relative depth information to be derived directly from horizontal pixel displacement magnitudes following equation 5.15. Recent learning-based architectures enable direct prediction of motion vectors, bypassing traditional optical flow formulation and optimization. The Recurrent All-Pairs Field Transforms (RAFT) network [293] was selected for vector field estimation due to its state-of-the-art performance, robust generalization capabilities, and efficient inference times. The network

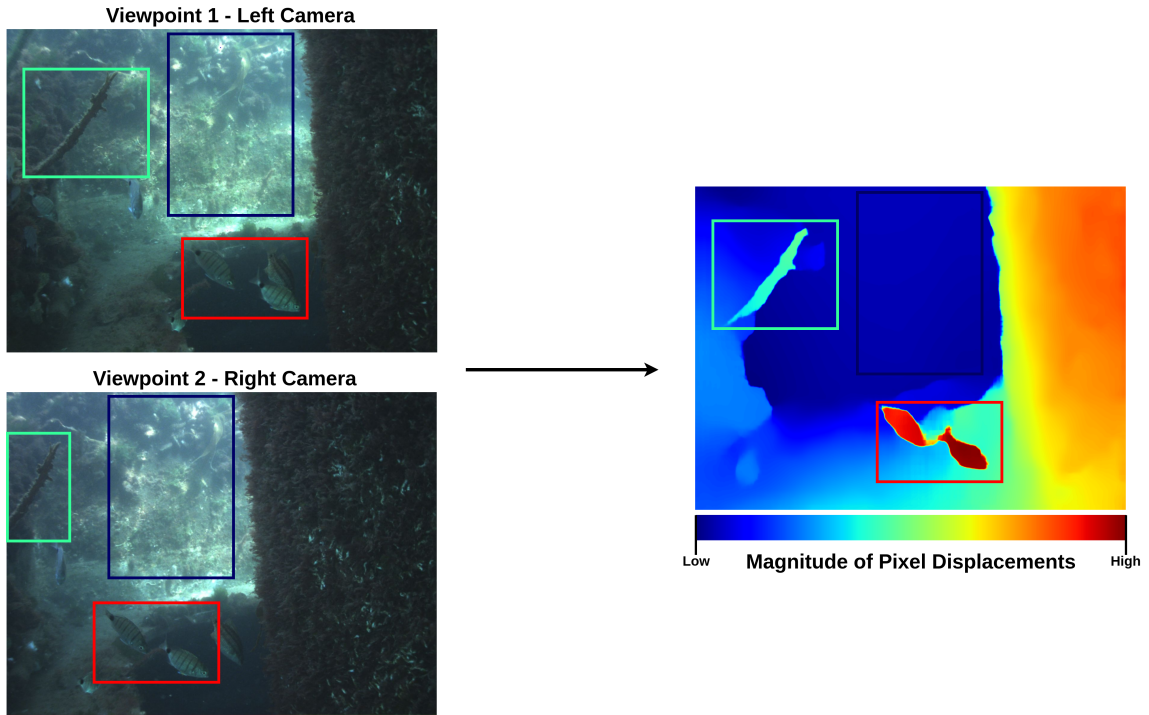


Figure 5.1: Depth information can be inferred from pixel displacements between two simultaneously captured viewpoints. The figure illustrates this principle: the fish (highlighted in red) are closer to the camera and therefore exhibit large pixel displacements between views, whereas the background wall (highlighted in blue) is more distant and shows minimal displacement, demonstrating how apparent motion magnitude inversely correlates with scene depth.

constructs multi-scale Four-Dimensional (4D) correlation volumes for all pixel pairs and iteratively refines flow predictions through a recurrent unit, ensuring accurate motion field estimation while maintaining a lightweight architecture suitable for real-time applications.

The NESO architecture follows an early fusion design to combine the magnitude of these pixel displacements with texture-based features from image data, to predict a dense depth representation of the scenario. NESO employs a pyramidal approach where the encoded high-level features are progressively reintroduced and merged during multiple stages of the upsampling process, allowing for better gradient flow and more detailed predictions. Figure 5.2 illustrates the NESO encoder-decoder learning-based model. The network processes a coupled input of dimensions (Channels $\times$ Height $\times$ Width) = $(4 \times 704 \times 1280)$, combining RGB features (texture, shadows, and occlusions) with pixel displacement magnitudes that encode relative depth information. Similarly to RHEA (section 4.3), feature extraction is performed through the lightweight ResNet backbone [284]. The ResNet architecture is adapted for multimodal fusion by modifying its initial convolutional layer to process a 4-channel input: three RGB components and one pixel displacement magnitude channel. The network is restructured for the depth estimation task by removing the original classification layers and introducing a 1×1 kernel 2D convolution that projects the feature maps to 1024 channels while preserving spatial dimensions. At this stage, the extracted features are compressed and encoded into low-resolution embeddings. The upsampling operation is conducted through

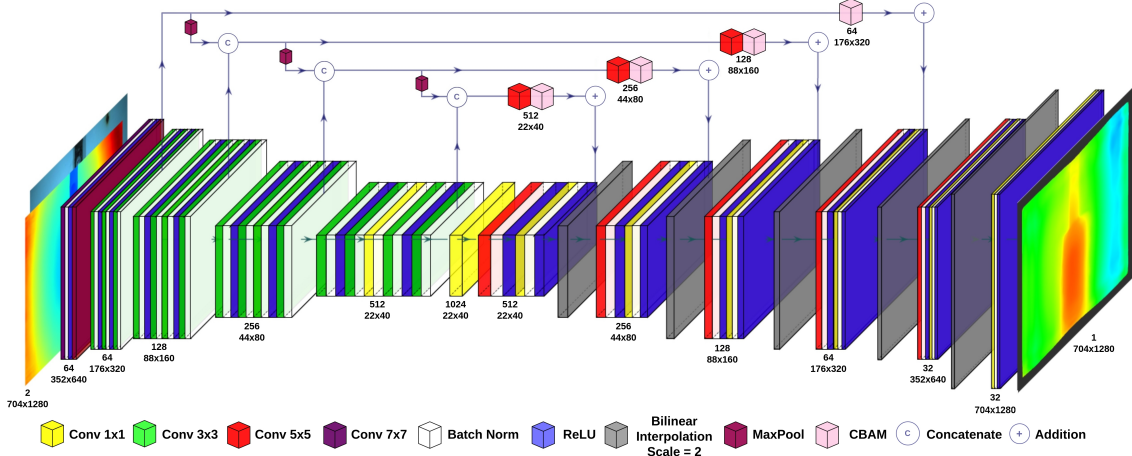


Figure 5.2: The NESO network follows an encoder–decoder architecture to combine RGB data with motion perception information to obtain a dense tri-dimensional representation of the underwater scenario. Pyramidal attention-based skip connections between both stages of the network promote that high-level features extracted during the encoding phase are reintroduced at multiple levels throughout the upsampling process. Feature maps are presented as Channels $\times$ Height $\times$ Width.

the BLConv5 decoder [42], discussed in section 4.3. While traditional skip connections connect encoder and decoder layers of matching dimensions [294], NESO adopts a more flexible approach by progressively incorporating high-level features from early network stages at different spatial resolutions during decoding. The feature maps are first asserted through maxpooling operations and concatenated channel-wise, followed by a 5×5 kernel depthwise convolution that merges the embeddings at each skip connection. The resulting representation is refined using Convolutional Block Attention Modules (CBAM) [295], which enhance feature discrimination through sequential channel and spatial attention mechanisms. This dual attention approach enables the network to focus on relevant feature channels while highlighting salient spatial regions, effectively suppressing noise and emphasizing discriminative features. The refined feature maps are then added to the corresponding decoder stage before the upsampling interpolation step. This pyramidal feature reintegration ensures robust gradient flow while preserving fine-grained spatial details, ultimately resulting in a high resolution depth prediction ($4 \times 704 \times 1280$).

The NESO network is trained and validated using the FLSea-Stereo dataset [205], which captures challenging underwater environments in the Mediterranean Sea characterized by high turbidity, homogeneous sandy regions, and marine growth-covered rocky outcrops. The dataset comprises 10987 training and 3663 validation stereo pairs, captured by two Nikon D810 1280×720 resolution cameras in watertight enclosures. Motion vectors are estimated for each stereo pair using the RAFT network. The pixel displacement magnitudes are normalized to obtain camera-independent depth cues, facilitating NESO’s learning task and generalization capabilities. Following equation 5.15, pixel displacements (Δx) depend on both camera parameters (focal length f_x and baseline B) and scene depth (P_z). Dividing these displacements by the optical system parameters ($-f_x b$) yields a normalized representation where disparity becomes directly proportional to inverse

depth. This normalization preserves relative depth relationships, where larger normalized values indicate closer objects and smaller values correspond to more distant elements in the scene, regardless of the imaging setup. Ground truth depth maps are generated using Agisoft Metashape's¹ SfM algorithm. The resulting 2.5D representation encodes metric depth through pixel intensity values, with valid measurements ranging from 0.5 to 38.0 meters. Readings outside this interval are considered filler pixels. While achieving errors consistently below 0.005 m for nearby objects, the reconstruction of distant elements exhibits imperfections due to underwater phenomena such as turbidity and backscattering.

Optimization of the network follows the same learning pattern used in section 4.3, accomplished through the custom loss function in equation 5.16:

$$l_{composed} = 0.45(l2_{valid}) + 0.10(l2_{filler}) + 0.45(l_{nbsmoothness}). \quad (5.16)$$

Equations 5.17, 5.18 and 5.19 further detail each loss term:

$$l2_{valid} = (i - \hat{i})^2, \quad \forall i \in [0.5, 38.0] \text{ m}, \quad (5.17)$$

$$l2_{filler} = (i - \hat{i})^2, \quad \forall i \notin [0.5, 38.0] \text{ m}, \quad (5.18)$$

$$l_{nbsmoothness} = \sum_{\Theta} ||e^{(-\nabla i)} \cdot \nabla \hat{i}||, \quad \forall i, \quad (5.19)$$

where i and $\hat{i}$ are the ground truth and estimated depth values, respectively; and Θ represents the pixel spatial domain.

The network optimization prioritizes valid pixels through an l2-based loss term, as these correspond directly to reliable ground truth depth measurements. A 10% penalty is attributed to the filler pixels that represent areas where ground truth data has been infilled. Scene consistency is regulated through the smoothness loss term (equation 5.19), which heavily penalizes depth discontinuities in high spatial frequency regions [296]. This term promotes faithful depth predictions particularly around edges and transition areas [297].

5.3 Results and Discussion

The content of this section aims to thoroughly discuss and analyze the performance of the fusion methodologies introduced along this chapter. The NESO network is first evaluated in the FLSea-Stereo dataset, where its results are compared to other relevant state-of-the-art architectures - refer to section 5.3.1. The generalization capabilities of the proposed network are then demonstrated within a controlled underwater environment (section 5.3.2). A decommissioned sacrificial anode is submerged within an indoor pool, while MARESy captures high-fidelity multimodal information

¹<https://www.agisoft.com/>

from the scene. An inspection procedure is then conducted by analyzing the dense point cloud obtained by using NESO to combine the retrieved RGB data and magnitude of pixel movement. The upsides of refining this representation through heterogeneous 3D data fusion principles are further discussed in this section, as the inspection procedure is repeated using JMR's tri-dimensional output.

The following metrics are used to evaluate output predictions and guide quantitative discussions:

- **MAE** - Mean Absolute Error.
- **REL** - Relative Error.
- **RMSE** - Root Mean Squared Error.
- δ_n - Percentage of Correctly Predicted Pixels.
- **GPU Time** - Network Elapsed Time.

The NESO network models are implemented in PyTorch and trained with the following hardware setup: NVIDIA GeForce RTX 2080 SUPER with 8 GB of VRAM, Intel i5-8600K CPU @ 3.60 GHz with 6 cores. SGD is used as optimization algorithm with a learning rate starting in 0.1, scheduled through an *1cycle* policy [298].

5.3.1 NESO State-of-the-Art Comparison

Diverting from state-of-the-art uni-modality monocular depth estimation networks, NESO combines texture-based features with the apparent movement of pixels to infer a dense tri-dimensional representation of the underwater scenario. This section studies the benefits of this approach, comparing its performance on the FLSea-Stereo dataset to other relevant learning-based methodologies: AdaBins [186], a transformer-based architecture that employs adaptive depth binning; ZoeDepth [187], which achieves robust cross-domain generalization through metric learning and a multi-head design; and UW_Depth [254], an encoder-decoder architecture with dense attention modules. The pre-trained models for AdaBins and ZoeDepth stem from monocular depth estimation tasks in aerial contexts, with proven cross-dataset generalization capabilities. The model used for UW_Depth is trained in the FLSea-VI dataset, which provides monocular data from the same underwater context as the dataset used to train the proposed NESO network. The FLSea-Stereo 0.5 to 38.0 m ground truth metric range is divided into several intervals and studied separately, allowing for a more detailed quantitative understanding of the faithfulness of each network's predictions.

The results depicted in table 5.1 reveal that NESO vastly exceeds the performance of other state-of-the-art networks, achieving an average RMSE of 0.771 m for the complete 0.5 to 38.0 m range. This metric error represents a 58.7% improvement over the UW_Depth network, that reaches the closest performance with an RMSE of 1.868 m. For the 0.5 to 2.0 m range, NESO

Table 5.1: State-of-the-Art comparison on the FLSea-Stereo dataset [205]. Available pre-trained models and default parameterization are used for state-of-the-art methods.

Method	Metric Range (m)	Modality	RMSE (m)	MAE (m)	δ_1 (%)	GPU Time (s)
AdaBins (2021) [186]	[0.5 - 38]	RGB	3.041	2.608	15.04	0.122
	[0.5 - 2]		3.121	3.093	0.00	
	[2 - 5]		2.126	1.962	18.03	
	[5 - 15]		2.724	2.102	44.53	
	[15 - 38]		13.018	12.754	0.00	
ZoeDepth (2023) [187]	[0.5 - 38]	RGB	2.267	1.559	23.59	0.265
	[0.5 - 2]		0.452	0.418	38.06	
	[2 - 5]		1.306	1.190	18.97	
	[5 - 15]		4.767	4.435	0.29	
	[15 - 38]		16.023	15.811	0.00	
Uw_Depth (2024) [254]	[0.5 - 38]	RGB	1.868	1.165	42.43	0.008
	[0.5 - 2]		0.674	0.591	42.87	
	[2 - 5]		1.005	0.847	55.08	
	[5 - 15]		3.466	3.030	14.16	
	[15 - 38]		13.813	13.547	0.00	
NESO	[0.5 - 38]	RGB + Mag. Pixel Displacements	0.771	0.468	78.04	0.029
	[0.5 - 2]		0.222	0.182	83.84	
	[2 - 5]		0.641	0.512	70.22	
	[5 - 15]		1.484	1.118	75.04	
	[15 - 38]		5.201	4.472	50.19	

Lower is better: RMSE, MAE, GPU Time

Higher is better: δ_1

outputs predictions with an average absolute error of 0.182 m and δ_1 of 83.84%, demonstrating its applicability for close-range inspection procedures. The AdaBins algorithm struggles the most in this metric range with 3.093 m of MAE, while the performance closest to NESO is achieved by ZoeDepth with an MAE and δ_1 of 0.418 m and 38.06%, respectively. Overall, table 5.1 depicts a steep performance decay for every network, as the distance range increases. This is to be expected, since the depth estimation task becomes vastly more difficult with distance, given that information retrieved by the observer becomes more scarce and compressed. Additionally, sub-sea phenomena such as absorption, marine snow and backscattering create a turbid and hazy environment that further hinders high-fidelity long-range data acquisition. Nevertheless, the depth map predictions estimated by NESO within the 15 to 38 m range have an average RMSE of 5.201 m - approximately 40% less than the closest state-of-the-art performance for this given range.

The UW_Depth network has the fastest GPU inference time with 0.008 s. AdaBins and ZoeDepth depend on heavier network backbones to extract features from each RGB input, and thus require longer processing times. The lightweight design of the NESO network allows for an inference time of 0.029 s. While it does not achieve the highest output prediction rate due to the more complex input representation, introducing motion perception principles as a network input proves a valuable addition, as NESO metrically outperforms uni-modality solutions.

A visual comparison of the predictions inferred by each network is presented in figure 5.3. The output estimated by NESO for Instance 1 shows a smooth depth gradient as the valley extends away from the observer. ZoeDepth and UW_Depth predict an accurate close-range representation, with the first capturing a more visually detailed prediction. As the distance grows, the error

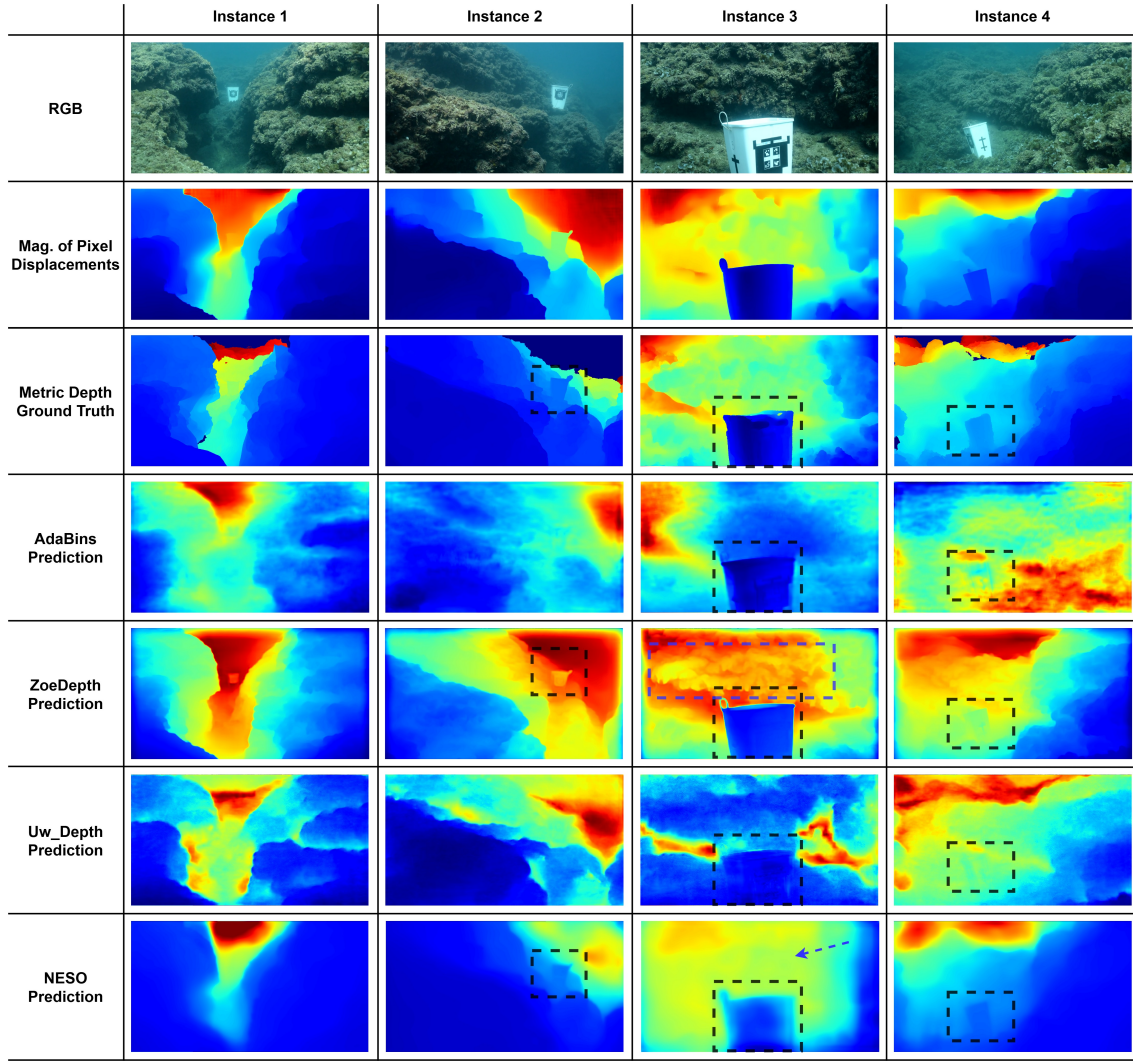


Figure 5.3: Qualitative comparison of depth predictions on FLSea-Stereo dataset samples. While AdaBins [186], ZoeDepth [187], and UW_Depth [254] rely solely on RGB input, NESO leverages both image data and pixel displacement magnitudes to enhance depth estimation.

in both estimations increases progressively as can be observed in the respective RMSE maps for Instance 1 - figure 5.4. The mid ranges prove to be the strong-suit of the AdaBins architecture. Figure 5.4 depicts this finding, as Instances 1, 2 and 4 clearly show that the error maps obtained by comparing AdaBins predictions to the respective ground truths have the least amount of error within this metric range. Table 5.1 further supports this analysis as the network achieves its best performance for the $[2, 5]$ and $[5, 15]$ m intervals, with mean absolute errors of 1.962 m and 2.102 m, respectively. The environment in Instance 2 of figure 5.3 illustrates how the scene becomes increasingly hazy and blurry with distance. Accurately representing the highlighted object, which lies just outside the camera's depth of field, is particularly challenging. Both AdaBins and UW_Depth fail to capture the object in their predictions. The haziness difficulties feature extraction, resulting in predictions where the object and background blend together. ZoeDepth offers

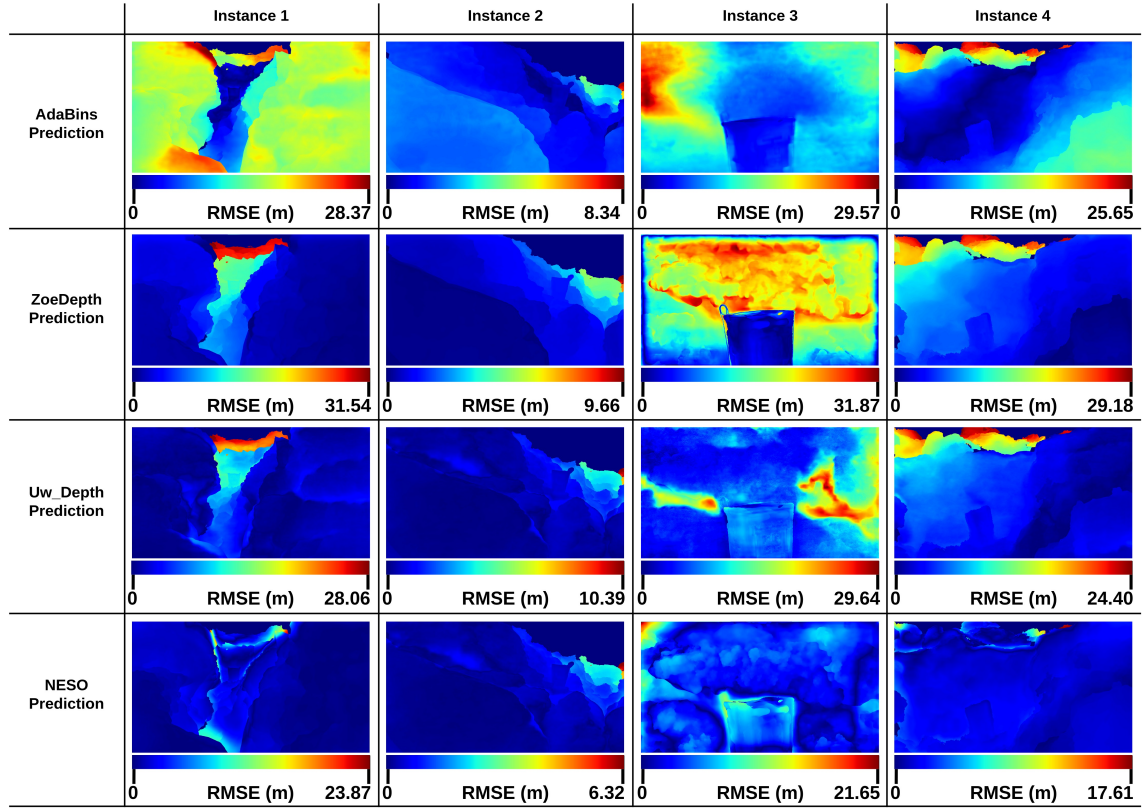


Figure 5.4: Root mean squared error maps comparing network predictions with ground truth depth measurements.

a more accurate visual representation of the object; however, introduces an offset error into the prediction that gets progressively more accentuated as the distance increases - refer to Instance 2, figure 5.4. The NESO network correctly represents the depth gradient of the scenario. Additionally, its pyramidal attention-based skip connections ensure that high-level features extracted from the encoder are reintroduced along multiple upsampling layers, allowing NESO to accurately depict the shape and depth of the highlighted object. This object is now the foreground focus of Instance 3 - figure 5.3. The UW_Depth network shows difficulty in separating the object from the background. While its prediction is metrically accurate overall, the high-depth gradient areas in the background account for most of the output's RMS error, as shown in figure 5.4. Instance 3 of figure 5.4 reveals that ZoeDepth estimates a very accurate foreground tri-dimensional representation. This result aligns with the findings depicted in table 5.1, as the network achieves an RMSE of 0.452 m for the [0.5, 2] m range. The output prediction inferred by NESO allows for an accurate tri-dimensional perception of the scene. Despite showing a high percentage of correctly predicted pixels in terms of δ_1 (averaging 78.04% across the complete metric range), the network presents indications of over-smoothing at the edges that outline the highlighted object, as well as in the background. Applying a stronger penalty to the $l_{nbsmoothness}$ loss term in equation 5.16 could encourage the network to focus more on these transitional areas. However, this increased emphasis on edge details is likely to degrade the network's overall metric performance for pixels

with valid depth. The current approach seeks a balance between neighborhood consistency and quantitative performance, hence the even weight penalty distribution. The final instance of figure 5.3 proves a difficult scenario. The state-of-the-art networks struggle with the increased level of haziness, leading to poor depth estimations past the foreground submerged structure on the right. NESO correctly depicts the depth gradient of the scene, and proves capable of representing the highlighted object, even in these harsh conditions. The additional input that provides the network with the apparent movement of pixels proves to be NESO's main advantage over other networks, since the relative depth cues inferred from motion perception principles offer complementary features when the texture-based input become less viable.

5.3.2 Controlled Underwater Experiments

Following the validation of NESO's depth estimation capabilities, this section examines the network's effectiveness in a controlled close-range inspection task, demonstrating its practical applicability for underwater structural assessment. The sacrificial anode used for this study exhibits a highly irregular surface due to its long operational lifespan. In the absence of original specifications or ground truth data, a set of relative measurements from key control features serves as baseline for evaluation, following the methodology presented along section 3.5.2. As illustrated in figure 3.15, several volumetric characteristics are assessed: peaks (most prominent surface areas), valleys (deepest eroded regions), anode wear (volumetric displacement between peaks and valleys), and distance to the metal core (measured from each peak to its corresponding extremity).

Evaluating NESO for Close-range Structural Inspection

A quantitative comparison of volumetric measurements between NESO and the best-performing baseline network (UW_Depth) is presented in table 5.2. Generalizing AI models to contexts outside their training and validation datasets can be challenging. However, the optimization scheme used during NESO's training facilitates its application to MARESyE data collected in close-range inspection scenarios. Table 5.2 reveals that NESO estimates the distance to the anode structure with offset absolute errors down to 0.1189 m and 0.1051 m, for the top and bottom sections of the anode, respectively. These errors sit well bellow the 0.182 m MAE expected for the FLSea-Stereo [0.5, 2] m range - refer to table 5.1 - proving that NESO can function for distinct scenarios and optical systems.

Relative measurements for peaks and valleys are derived from the metric difference in depth values between each control feature and the background pool wall. Figure 5.5 portrays the depth predictions inferred by both the NESO and UW_Depth networks for the top and bottom sections of the anode structure. The state-of-the-art UW_Depth architecture estimates an inaccurate and irregular representation of the background wall. Additionally, the RGB uni-modality network struggles to handle the anode's irregular surface topology, showing inconsistent depth estimation across different features. While valley regions are estimated with millimetric accuracy (absolute errors of 0.0075 m and 0.0098 m for top and bottom sections), peak measurements demonstrate significantly higher uncertainty with relative errors reaching 51.82%. NESO outputs depth map

Table 5.2: Volumetric assessment of the anode reconstructions estimated by both the NESO and UW_Depth networks.

Characteristics		Ground Truth	Uw_Depth [254]			NESO		
		Measured (m)	Estimated (m)	Absolute Error (m)	Relative Error (%)	Estimated (m)	Absolute Error (m)	Relative Error (%)
Top	Peak Depth	0.0793	0.0960	0.0167	21.06	0.0913	0.0120	15.13
	Valley Depth	0.0723	0.0648	0.0075	10.37	0.0836	0.0113	15.62
	Anode Wear	0.0070	0.0312	0.0242	345.71	0.0077	0.0007	10.00
	Distance to Core	0.0350	0.0097	0.0253	72.29	0.0334	0.0016	4.57
	Distance to Observer	0.4100	0.2651	0.1449	35.34	0.2911	0.1189	29.00
Bottom	Peak Depth	0.0768	0.1166	0.0398	51.82	0.0496	0.0272	35.42
	Valley Depth	0.0729	0.0827	0.0098	13.44	0.0433	0.0296	40.60
	Anode Wear	0.0039	0.0338	0.0299	766.67	0.0063	0.0024	61.54
	Distance to Core	0.0350	0.0327	0.0023	6.57	0.0247	0.0103	29.43
	Distance to Observer	0.4230	0.2198	0.2032	48.04	0.3179	0.1051	24.85

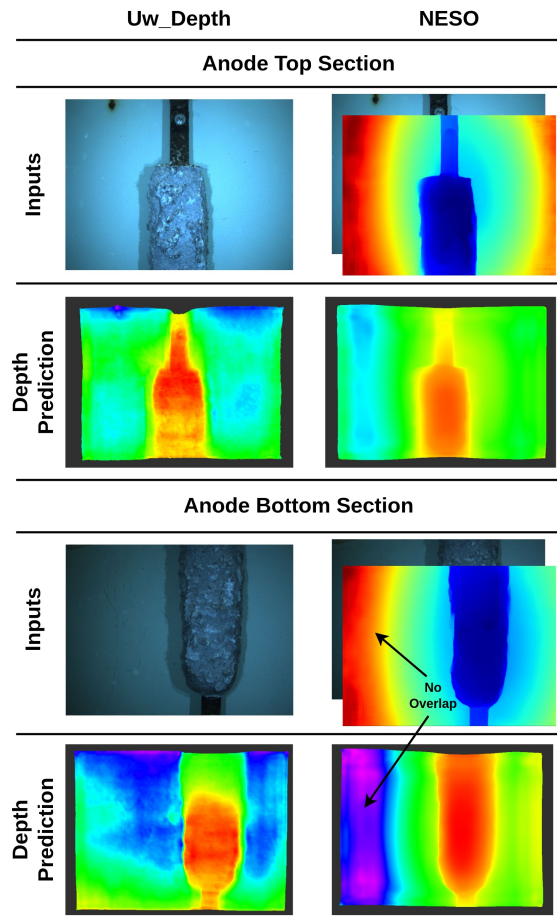


Figure 5.5: NESO and UW_Depth [254] depth map predictions for the top and bottom sections of a sacrificial anode, inferred during a close-range inspection procedure conducted within controlled underwater environments.

predictions that smoothly represent the pool background, except for the left portion of the image (refer to figure 5.5). This is due to MARESyE's stereo image pair lack of overlap in this region: the RAFT network can't find pixel correspondences in between the two viewpoints for this area,

leading to invalid magnitude values. These pixels are filtered out, and not considered during the anode's assessment. The volumetric representation estimated by NESO assesses the top section of the anode to have a peak of 0.0960 m and a valley of 0.0836 m; corresponding to average relative errors of approximately 15%. Both networks estimate measurements with higher degrees of error for the bottom half of the anode, due to the challenging conditions caused by the darker illumination of the scene. For features related to the structural integrity of the anode, NESO outperforms UW_Depth, achieving millimeter-precise predictions. The anode wear, as measured from NESO's depth map, shows absolute errors as low as 0.0007 m and 0.0024 m for the top and bottom sections, respectively. In contrast, UW_Depth struggles to capture these intricate details, estimating anode wear with absolute errors exceeding 0.02 m. NESO's prediction of the distance from the top portion of the anode to its core results in an absolute error of 0.0016 m, corresponding to a relative error of 4.57%. However, the same level of precision is not achieved for the lower section of the anode, where more challenging conditions and over-smoothing in certain point cloud regions lead to higher absolute errors.

The NESO network demonstrates its ability to capture features directly from the coupled RGB/magnitude of pixel displacements input, without requiring extensive parameterization, to infer dense and precise tri-dimensional representations of the target sacrificial anode.

Enhancing NESO Predictions through Active Imaging Integration

A thorough inspection of submerged structures, such as sacrificial anodes, requires millimeter-precise information. While NESO point clouds provide accurate representations of the underwater environment, particularly in conveying structural integrity, its metric precision can be enhanced when operating outside the network's training domain. This limitation stems from NESO's training on the FLSea-Stereo dataset, which spans a wide metric range (0.5 to 38 m) with sharp depth gradients, differing significantly from MARESy's optimal inspection workspace of 0.5 to 0.6 m. The integration of millimeter-precise LSR measurements offers a complementary solution to this challenge, as the active sensing data can refine NESO's predictions within its operating range. By fusing these complementary modalities, the system leverages both the dense coverage of NESO's predictions and the high precision of LSR measurements, enabling more accurate metric assessment of structural features during close-range inspection tasks. The following required packages to your document preamble:

The inspection of the target anode demonstrates significant benefits from fusing NESO predictions with sparse LSR data. As shown in table 5.3, the JMR algorithm achieves notable improvements in metric accuracy, particularly in estimating the distance from the observer, with gains exceeding 0.10 m for both top and bottom sections. The fused representation achieves absolute errors as low as 0.0065 m and 0.0033 m, corresponding to relative errors of 1.59% and 0.78% for each section, respectively. Figure 5.6 illustrates the fusion process, comparing NESO's initial prediction, LSR measurements, and the final refined point cloud. MARESy's strategic laser beam positioning, reflecting one beam off the anode and another off the pool wall, ensures optimal scene coverage. The JMR algorithm identifies two distinct homogeneous regions - anode and

Table 5.3: Volumetric assessment of the anode reconstruction obtained by fusing NESO's prediction with sparse LSR data through the JMR algorithm.

Characteristics		Ground Truth	NESO - JMR			Improvement over NESO (m)
		Measured (m)	Estimated (m)	Absolute Error (m)	Relative Error (%)	
Top	Peak Depth	0.0793	0.0816	0.0023	2.90	0.0097
	Valley Depth	0.0723	0.0726	0.0003	0.41	0.0110
	Anode Wear	0.0070	0.0090	0.0020	28.57	-0.0013
	Distance to Core	0.0350	0.0384	0.0034	9.71	-0.0018
	Distance to Obs.	0.4100	0.4165	0.0065	1.59	0.1124
Bottom	Peak Depth	0.0768	0.0693	0.0075	9.77	0.0197
	Valley Depth	0.0729	0.0627	0.0102	13.99	0.0194
	Anode Wear	0.0039	0.0065	0.0026	66.67	-0.0002
	Distance to Core	0.0350	0.0196	0.0154	44.00	-0.0051
	Distance to Obs.	0.4230	0.4263	0.0033	0.78	0.1018

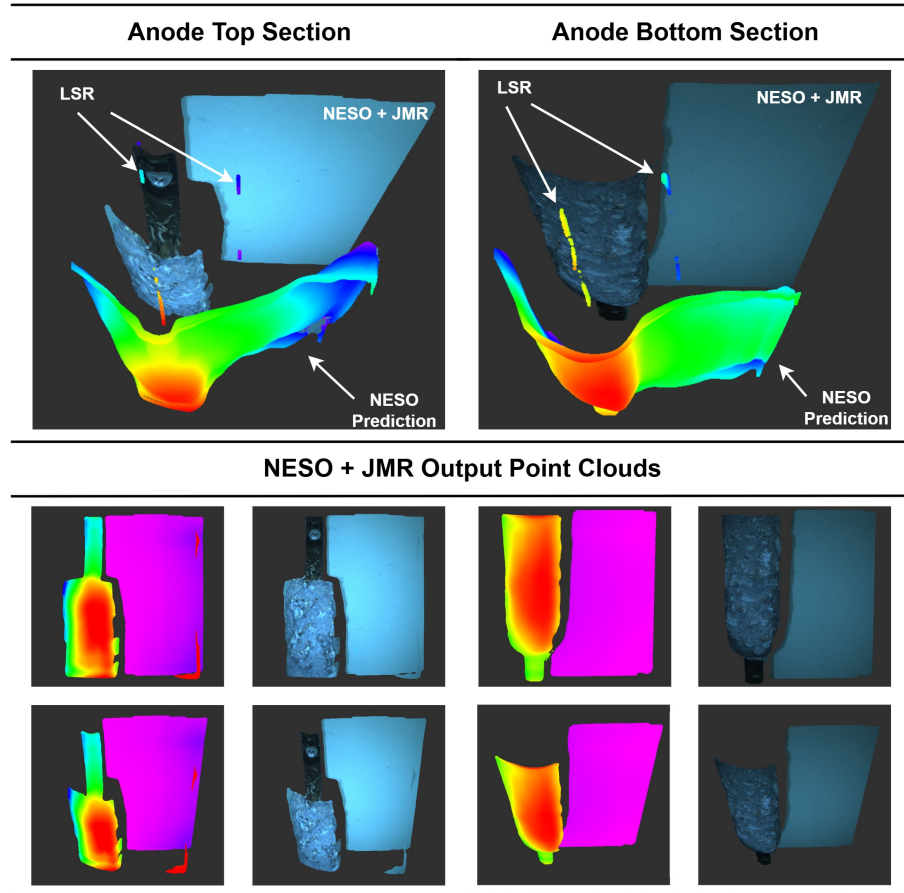


Figure 5.6: Sacrificial anode reconstructions obtained by fusing sparse LSR information with the dense NESO point cloud prediction, using the JMR algorithm.

background - applying separate regression curves to map NESO-LSR depth displacements in each area. This approach effectively eliminates over-smoothing at transition regions while maintaining

the high density of NESO's predictions. The fusion process shows particular strength in volumetric assessment, achieving improvements of 0.0197 m and 0.0194 m for peak and valley features in the bottom section, with top region measurements reaching relative errors as low as 0.41%. However, the JMR algorithm exhibits some limitations in handling curved surfaces, where peripheral points deviating from the laser beam guidance tend to be over-corrected [38]. This results in a slight performance decline, with increased absolute errors up to 0.0051 m for certain structural features. Despite these minor limitations, the quantitative results validate the effectiveness of the fusion approach in achieving both high point density and millimeter-precise measurements.

5.4 Final Considerations

This chapter introduces a comprehensive approach to underwater scene understanding through multimodal data fusion, addressing the inherent limitations of individual sensing modalities: image data can only provide a 2D contextual perspective of the scene, apparent motion-based components are limited to a relative sense of depth, and active imaging techniques provide millimeter-precise measurements with constrained spatial coverage. The proposed NESO encoder-decoder network leverages these complementary characteristics by fusing textured data with motion perception principles to create dense 3D scene representations. Its attention-based pyramidal architecture ensures that relevant features from the coupled input propagate through multiple decoding stages, preserving crucial information for depth estimation. This approach surpasses other depth estimation models by 58.7% in RMSE on the FLSea-Stereo dataset, while showing strong generalization capabilities in reconstructing a submerged sacrificial anode with absolute errors as low as 0.0016 m. Despite providing accurate relative measurements, the dense NESO prediction struggles with absolute depth values. Heterogeneous tri-dimensional fusion principles are employed to further refine the high-density point cloud, leveraging millimeter-precise active imaging data. A regression-based algorithm maps depth displacements between these heterogeneous 3D point clouds, combining the density and contextual richness of NESO predictions with the metric precision of laser triangulation, resulting in metric improvements of up to 0.1124 m. These advances in underwater imaging enhance the perceptual capabilities of underwater vehicles, enabling more reliable inspection and maintenance operations by fusing complementary sensing modalities into an unified tri-dimensional representation of the sub-sea environment.

Chapter 6

Conclusion

Current O&M methodologies rely heavily on manual operations, requiring divers, large vessels, support personnel, and specialized equipment. Despite strict safety regulations, human operators remain exposed to significant risks, while weather limitations restrict the yearly operational window. Unmanned underwater vehicles offer safer alternatives, but their widespread adoption is limited by inadequate perception capabilities. Traditional solutions using vessel-mounted sensors can lead to occlusions and increase risk to the vehicle when approaching structures for detailed observations. Eye-in-hand configurations, where sensors are coupled directly to robotic manipulators, can enhance situational awareness by leveraging the arm's degrees of freedom to inspect from previously inaccessible angles. However, conventional underwater perception systems are unsuitable for these configurations and suffer severe data degradation from sub-sea phenomena. The lack of reliable, high-fidelity tri-dimensional information fundamentally limits the ability to perform autonomous close-range inspection and manipulation tasks, hindering the progression toward fully autonomous offshore O&M operations.

This research addressed these fundamental perception limitations through an integrated approach combining sensor design with data fusion methodologies. The MARESyE hybrid imaging sensor employed mitigation strategies to reduce the degradation of the gathered multimodal information. Distinct data streams provide complementary characteristics but exhibit inherent error distributions and uncertainties. Information fusion principles were introduced to leverage the strengths of each complementary modality, integrating them into a unified tri-dimensional representation of the environment. Heterogeneous 3D fusion was achieved through both the Joint Masked Regression (JMR) algorithm and the underWater HEterogenous 3d information fusion Architecture (RHEA) to combine point clouds from different imaging techniques. Additionally, the fusiNg tEXture with apparent motion information for underwater Scene recOnstruction (NESO) architecture introduced a learning-based approach that fuses textured images with apparent motion into dense tri-dimensional representations.

The MARESyE hybrid imaging sensor was developed to advance the state of the art of underwater perception by gathering high-fidelity data from sub-sea environments. Diverging from traditional approaches that rely on a single source of information, MARESyE integrates both pas-

sive stereo imaging and active laser projection. A data-driven range-gated scheme synchronizes the acquisition system to receive light reflections from the scenario, while avoiding backscattered photons from nearby suspended particles. The optical design employs hemispherical dome ports to minimize light refraction effects, while the careful positioning of cameras, lasers, and external light sources enables high-quality close-range data capture within a compact housing depth-rated for 200 m. Thorough experimental validation demonstrated millimeter-precise measurements within the optimal 0.5-0.6 m operating range. A comprehensive performance characterization revealed that both photogrammetric stereo and light stripe ranging techniques achieve sub-millimeter accuracy in global depth estimation for close ranges, though LSR maintains consistent performance over longer distances. For relative depth measurements, the stereo-based methodology proved more stable, containing the mean absolute error within 0.01 m, while the LSR technique showed progressively degraded performance for longer ranges due to light absorption and reduced laser-camera resolution. Controlled tests with a decommissioned sacrificial anode validated these findings, where MARESy accurately captured its irregular surface topology, estimating peaks, valleys and wear patterns with absolute errors down to 0.0008 m. Real-world validation at the ATLANTIS Coastal Testbed confirms the sensor's robustness in harsh maritime conditions, maintaining reliable performance despite the additional challenges posed by biofouling, varying light conditions, and wave-induced motion.

Information fusion in underwater environments is crucial for mitigating the inherent error characteristics of individual techniques or sensors. Measurements from each source are subject to uncertainty, whether from noise introduced by the environment or the sensor itself. Passive stereo imaging generates dense point clouds with broad scene coverage but suffers from matching ambiguities in low-texture regions and degraded accuracy beyond calibration range, while active laser triangulation achieves millimeter-precise depth measurements but provides sparse data with limited spatial coverage. Novel heterogeneous 3D fusion methodologies were developed to leverage these complementary strengths while mitigating individual limitations. The Joint Masked Regression algorithm exploits the fundamental principle that homogeneous regions around laser projections generally translate the same depth. By propagating accurate laser measurements across these regions through regression curves, JMR maintained the density of stereo-based data while correcting it with precise laser information. The RHEA network employs an encoder-decoder architecture to fuse coupled 3D inputs through an early fusion approach. A lightweight residual encoder extracts features directly from both point clouds while maintaining computational efficiency for edge processing. The network was trained through a synthetic-to-real scheme that carefully models the characteristics of data recovered from each underwater imaging technique, enabling direct deployment without requiring domain adaptation. Extensive validation in controlled underwater trials demonstrated the effectiveness of both approaches, with JMR and RHEA achieving average absolute errors of 0.0133 m and 0.0097 m respectively, corresponding to improvements of 34.9% and 52.5% over baseline measurements. Experiments at the ATLANTIS Test Center validated these methodologies in real conditions, where both approaches successfully enhanced point cloud accuracy despite the heightened effects of sub-sea phenomena. In essence, JMR's ef-

fectiveness was constrained by laser coverage and positioning, as the algorithm requires sufficient resolution to generate faithful regression curves, while RHEA demonstrated greater flexibility with sparse inputs at the cost of occasional artifacts in edge regions and transition areas. Despite these advances in heterogeneous 3D fusion, fundamental photogrammetric limitations and extensive parameterization still constrain accuracy. To address these constraints, a novel multimodal approach was developed through the NESO architecture. The network leverages the inverse relationship between apparent pixel motion magnitude and scene depth, combining textured data with motion perception principles through an attention-based pyramidal design. This approach allows high-level features to propagate through multiple decoding stages, preserving crucial scene information while minimizing the need for extensive parameter tuning. Evaluation on the FLSea-Stereo dataset demonstrated NESO's effectiveness, achieving a 58.7% RMSE improvement over uni-modality state-of-the-art networks. While NESO generated dense and textured point clouds with accurate relative depth relationships, its predictions exhibited a global depth offset when operating outside its training metric range. This limitation was addressed by fusing the dense predictions with sparse laser data through the JMR algorithm, achieving absolute metric improvements of up to 0.1124 m.

The scientific outcomes of this research advanced underwater perception capabilities by combining data-driven imaging techniques with novel information fusion methodologies. The compact design of MARESyE makes it particularly suitable for eye-in-hand configurations, enabling close-range inspection from multiple viewpoints without compromising vehicle safety. MARESyE's architecture fostered the implementation of both heterogeneous 3D and multimodal fusion principles by ensuring precise temporal synchronization of data streams and accurate spatial alignment between imaging modalities. These technological advances, coupled with embedded processing capabilities, enable the real-time deployment of fusion algorithms that integrate complementary information into unified scene representations. These developments significantly enhanced three-dimensional scene understanding in harsh sub-sea environments, providing the high-fidelity perception capabilities required for autonomous offshore inspection and maintenance operations.

6.1 Future Work

Several promising research directions emerge from this work. The integration of advanced image enhancement methodologies could further improve the quality of gathered information, where model-based restoration combined with bilateral filtering and dark channel prior techniques could enhance contrast and reduce haze. These preprocessing pipelines would need to be optimized for real-time operation to maintain edge processing capabilities. Future developments of the MARESyE sensor focus on a continuous improvement of its data acquisition process. Employing more powerful laser beam projectors can help mitigate the underwater absorption phenomena and ensure these are visible enough to be usable within the LSR module. Furthermore, deep learning-based stereo algorithms can be explored to replace the currently used SGBM procedure. Fusing the heterogeneous 3D information directly in point cloud form through the use of 3D CNNs or Graph Neural Networks can also be a valid solution, bypassing the need to project the information

into a 2D plane. A more long term hypothesis is developing an end-to-end architecture that learns to extract tri-dimensional data directly from the captured 2D images and fuses both modules internally. Future research can also focus on developing a modular version of MARESyE based on the insights gained from initial eye-in-hand field deployments. The additional degrees of freedom provided by robotic manipulators can be exploited to gather complementary information from distinct viewpoints, enabling multi-view fusion approaches to effectively combine data gathered from different arm poses, requiring robust registration techniques and uncertainty-aware fusion strategies to handle the cumulative error from sequential acquisitions. The final demonstration these developments should occur through comprehensive trials in operational offshore wind farms, where the complete AI-driven eye-in-hand MARESyE system can be tested under real conditions.

References

- [1] Intergovernmental Panel on Climate Change (IPCC). *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*. Cambridge University Press, 2022.
- [2] Christoph Kern and Andreas Jess. Reducing global greenhouse gas emissions to meet climate targets—a comprehensive quantification and reasonable options. *Energies*, 14(17), 2021. doi:10.3390/en14175260.
- [3] Xuefei Wang, Xiangwu Zeng, Jiale Li, Xu Yang, and Haijun Wang. A review on recent advancements of substructures for offshore wind turbines. *Energy Conversion and Management*, 158:103–119, 2018. doi:10.1016/j.enconman.2017.12.061.
- [4] Giuseppe Costanzo, Guy Brindley, Guy Willems, Lizet Ramirez, Phil Cole, and Vasiliki Klonari. Wind energy in europe: 2023 statistics and the outlook for 2024-2030. Technical report, WindEurope, 2024.
- [5] Saravanakumar Rajendran, Matias Diaz, Roberto Cárdenas, Enrique Espina, Emilio Contreras, and Jose Rodriguez. A review of generators and power converters for multi-mw wind energy conversion systems. *Processes*, 10(11), 2022. doi:10.3390/pr10112302.
- [6] Andry Maykol Pinto, João V. Amorim Marques, Daniel Filipe Campos, Nuno Abreu, Aníbal Matos, Martio Jussi, Robin Berglund, Jari Halme, Petri Tikka, João Formiga, et al. Atlantis-the atlantic testing platform for maritime robotics. In *OCEANS 2021: San Diego–Porto*, pages 1–5. IEEE, 2021. doi:10.23919/OCEANS44145.2021.9706059.
- [7] Zhengru Ren, Amrit Shankar Verma, Ye Li, Julie J.E. Teuwen, and Zhiyu Jiang. Offshore wind turbine operations and maintenance: A state-of-the-art review. *Renewable and Sustainable Energy Reviews*, 144:110886, 2021. doi:10.1016/j.rser.2021.110886.
- [8] Dominique Roddier, Christian Cermelli, Alexia Aubault, and Alla Weinstein. Windfloat: A floating foundation for offshore wind turbines. *Journal of Renewable and Sustainable Energy*, 2(3):033104, 2010. doi:10.1063/1.3435339.
- [9] Anders Myhr, Catho Bjerkseter, Anders Ågotnes, and Tor A. Nygaard. Levelised cost of energy for offshore floating wind turbines in a life-cycle perspective. *Renewable Energy*, 66:714–728, 2014. doi:10.1016/j.renene.2014.01.017.
- [10] Brian Snyder and Mark J. Kaiser. Ecological and economic cost-benefit analysis of offshore wind energy. *Renewable Energy*, 34(6):1567–1578, 2009. doi:10.1016/j.renene.2008.11.015.

- [11] Ying Du, Shengxi Zhou, Xingjian Jing, Yeping Peng, Hongkun Wu, and Ngaiming Kwok. Damage detection techniques for wind turbine blades: A review. *Mechanical Systems and Signal Processing*, 141:106445, 2020. doi:10.1016/j.ymssp.2019.106445.
- [12] John V. Sharp and Gerhard Ersdal. Damage types in offshore structures. In *Underwater Inspection and Repair for Offshore Structures*, chapter 3, pages 37–78. John Wiley & Sons Ltd, 2021. doi:10.1002/9781119633891.ch3.
- [13] Omer Khalid, Guangbo Hao, Cian Desmond, Hamish Macdonald, Fiona Devoy McAuliffe, Gerard Dooly, and Weifei Hu. Applications of robotics in floating offshore wind farm operations and maintenance: Literature review and trends. *Wind Energy*, 25(11):1880–1899, 2022. doi:10.1002/we.2773.
- [14] Daniel Filipe Barros Campos. *Multi-domain Contextual Awareness using Unmanned Surface Vehicles for Offshore Wind Farms Inspection*. PhD thesis, Faculty of Engineering of the University of Porto, FEUP, 2024. URL: <https://hdl.handle.net/10216/161860>.
- [15] Sara Bernardini, Ferdian Jovan, Zhengyi Jiang, Simon Watson, Andrew Weightman, Peiman Moradi, Tom Richardson, Rasoul Sadeghian, and Sina Sareh. A multi-robot platform for the autonomous operation and maintenance of offshore wind farms. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, page 1696–1700, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. doi:10.5555/3398761.3398956.
- [16] Ferdian Jovan and Sara Bernardini. Adaptive temporal planning for multi-robot systems in operations and maintenance of offshore wind farms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15782–15788, 2024. doi:10.1609/aaai.v37i13.26874.
- [17] Jean Goyet, Daniel Straub, and Michael H. Faber. Risk-based inspection planning of offshore installations. *Structural Engineering International*, 12(3):200–208, 2002. doi:10.2749/101686602777965360.
- [18] Harriet Fox, Ajit C. Pillai, Daniel Friedrich, Maurizio Collu, Tariq Dawood, and Lars Johanning. A review of predictive and prescriptive offshore wind farm operation and maintenance. *Energies*, 15(2), 2022. doi:10.3390/en15020504.
- [19] Helen Hastie, Katrin Lohan, Mike Chantler, David A. Robb, Subramanian Ramamoorthy, Ron Petrick, Sethu Vijayakumar, and David Lane. The orca hub: Explainable offshore robotics through intelligent interfaces, 2018. doi:10.48550/arXiv.1803.02100.
- [20] Maricruz Fun Cepeda, Marcos Freitas Machado, Fabrício Hudson Barbosa, Douglas Souza Moreira, Maria José Almansa, Marcelo Igor de Souza, and Jean-David Caprace. Exploring autonomous and remotely operated vehicles in offshore structure inspections. *Journal of Marine Science and Engineering*, 11(11), 2023. doi:10.3390/jmse11112172.
- [21] Romano Capocci, Gerard Dooly, Edin Omerdić, Joseph Coleman, Thomas Newe, and Daniel Toal. Inspection-class remotely operated vehicles—a review. *Journal of Marine Science and Engineering*, 5(1), 2017. doi:10.3390/jmse5010013.
- [22] Chenyu Zhao, Philipp R. Thies, and Lars Johanning. Offshore inspection mission modelling for an asv/rov system. *Ocean Engineering*, 259:111899, 2022. doi:10.1016/j.oceaneng.2022.111899.

- [23] Giorgio Ioannou, Nicola Forti, Leonardo M. Millefiori, Sandro Carniel, Alfredo Renga, Giuseppe Tomasicchio, Stefano Binda, and Paolo Braca. Underwater inspection and monitoring: Technologies for autonomous operations. *IEEE Aerospace and Electronic Systems Magazine*, 39(5):4–16, 2024. doi:10.1109/MAES.2024.3366144.
- [24] Xianbo Xiang, Caoyang Yu, Zemin Niu, and Qin Zhang. Subsea cable tracking by autonomous underwater vehicle with magnetic sensing guidance. *Sensors*, 16(8), 2016. doi:10.3390/s16081335.
- [25] Mohammed E. Sayed, Markus P. Nemitz, Simona Aracri, Alistair C. McConnell, Ross M. McKenzie, and Adam A. Stokes. The limpet: A ros-enabled multi-sensing platform for the orca hub. *Sensors*, 18(10), 2018. doi:10.3390/s18103487.
- [26] Roger Pi, Patryk Cieślak, Joan Esteba, Narcís Palomeras, and Pere Ridao. Compliant manipulation with quasi-rigid docking for underwater structure inspection. *IEEE Access*, 11:128957–128969, 2023. doi:10.1109/ACCESS.2023.3332486.
- [27] Dinh Quang Huy, Nicholas Sadjoli, Abu Bakr Azam, Basman Elhadidi, Yiyu Cai, and Gerald Seet. Object perception in underwater environments: a survey on sensors and sensing methodologies. *Ocean Engineering*, 267:113202, 2023. doi:10.1016/j.oceaneng.2022.113202.
- [28] Franka Nauert and Peter Kampmann. Inspection and maintenance of industrial infrastructure with autonomous underwater robots. *Frontiers in Robotics and AI*, 10, 2023. doi:10.3389/frobt.2023.1240276.
- [29] Jeremy Dillon and Richard Charron. Resolution measurement for synthetic aperture sonar. In *OCEANS 2019 MTS/IEEE SEATTLE*, pages 1–6, 2019. doi:10.23919/OCEANS40490.2019.8962823.
- [30] Andry Maykol Pinto, Henrique Pinto, and Anibal C. Matos. A mosaicking approach for visual mapping of large-scale environments. In *2016 International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 87–93, 2016. doi:10.1109/ICARSC.2016.18.
- [31] Pedro Nuno Leite, Pedro Nuno Pereira, João M.M. Dionísio, and Andry Maykol Pinto. Hybrid underwater imaging for the tri-dimensional inspection of critical structural elements in offshore platforms. *Ocean Engineering*, 314:119658, 2024. doi:10.1016/j.oceaneng.2024.119658.
- [32] Sara Aldhaferi, Giulia De Masi, Èric Pairet, and Paola Ardón. Underwater robot manipulation: Advances, challenges and prospective ventures. In *OCEANS 2022 - Chennai*, pages 1–7, 2022. doi:10.1109/OCEANSSChennai45887.2022.9775489.
- [33] Kerry He, Rhys Newbury, Tin Tran, Jesse Haviland, Ben Burgess-Limerick, Dana Kulić, Peter Corke, and Akansel Cosgun. Visibility maximization controller for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(3):8479–8486, 2022. doi:10.1109/LRA.2022.3188430.
- [34] Hai Huang, Xinyu Bian, Fengchun Cai, Jiyong Li, Tao Jiang, Zhenkun Zhang, and Chaoyu Sun. A review on visual servoing for underwater vehicle manipulation systems automatic control and case study. *Ocean Engineering*, 260:112065, 2022. doi:10.1016/j.oceaneng.2022.112065.

- [35] P. J. Sanz, A. Peñalver, J. Sales, D. Fornas, J. J. Fernández, J. Pérez, and J. Bernabé. Grasper: A multisensory based manipulation system for underwater operations. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 4036–4041, 2013. doi:10.1109/SMC.2013.689.
- [36] Giacomo Marani, Song K. Choi, and Junku Yuh. Underwater autonomous manipulation for intervention missions AUVs. *Ocean Engineering*, 36(1):15–23, 2009. doi:10.1016/j.oceaneng.2008.08.007.
- [37] Antonio Peñalver, J. Javier Fernandez, Jorge Sales, and Pedro J. Sanz. Multi-view underwater 3D reconstruction using a stripe laser light and an eye-in-hand camera. *MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World*, 2015. doi:10.1109/OCEANS-Genova.2015.7271497.
- [38] Pedro Nuno Leite and Andry Maykol Pinto. Fusing heterogeneous tri-dimensional information for reconstructing submerged structures in harsh sub-sea environments. *Information Fusion*, 103:102126, 2024. doi:10.1016/j.inffus.2023.102126.
- [39] Bahador Khaleghi, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013. doi:10.1016/j.inffus.2011.08.001.
- [40] Su-Fang Zhang, Jun-Hai Zhai, Bo-Jun Xie, Yan Zhan, and Xin Wang. Multi-modal representation learning: Advances, trends and challenges. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6, 2019. doi:10.1109/ICMLC48188.2019.8949228.
- [41] Andry Maykol Pinto, João V. Amorim Marques, Nuno Abreu, Daniel Filipe Campos, Maria Inês Pereira, Eduardo Gonçalves, Hugo Jorge Campos, Pedro Pereira, Francisco Neves, Aníbal Matos, Shashank Govindaraj, and Lillian Durand. Atlantis coastal testbed: A near-real playground for the testing and validation of robotics for o&m. In *OCEANS 2023 - Limerick*, pages 1–5, 2023. doi:10.1109/OCEANSLimerick52467.2023.10244595.
- [42] Pedro Nuno Leite and Andry Maykol Pinto. Exploiting motion perception in depth estimation through a lightweight convolutional neural network. *IEEE Access*, 9:76056–76068, 2021. doi:10.1109/ACCESS.2021.3082697.
- [43] Pedro Nuno Leite, Renato Jorge Silva, Daniel Filipe Campos, and Andry Maykol Pinto. Dense disparity maps from rgb and sparse depth information using deep regression models. In Aurélio Campilho, Fakhri Karray, and Zhou Wang, editors, *Image Analysis and Recognition*, pages 379–392, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-50347-5_33.
- [44] João M. M. Dionísio, Pedro N. A. A. S. Pereira, Pedro N. Leite, Francisco S. Neves, João Manuel R. S. Tavares, and Andry M. Pinto. Nereon - an underwater dataset for monocular depth estimation. In *OCEANS 2023 - Limerick*, pages 1–7, 2023. doi:10.1109/OCEANSLimerick52467.2023.10244675.
- [45] João Mina, Pedro Nuno Leite, João Carvalho, Lourenço Pinho, Eduardo P. Gonçalves, and Andry Maykol Pinto. Enhancing underwater inspection capabilities: A learning-based approach for automated pipeline visibility assessment. In Lino Marques, Cristina Santos, José Luís Lima, Danilo Tardioli, and Manuel Ferre, editors, *Robot 2023: Sixth*

- Iberian Robotics Conference*, pages 162–173, Cham, 2024. Springer Nature Switzerland. doi:10.1007/978-3-031-59167-9_14.
- [46] João Carvalho, Pedro Nuno Leite, João Mina, Lourenço Pinho, Eduardo P. Gonçalves, and Andry Maykol Pinto. Artificial intelligence for automated marine growth segmentation. In Lino Marques, Cristina Santos, José Luís Lima, Danilo Tardioli, and Manuel Ferre, editors, *Robot 2023: Sixth Iberian Robotics Conference*, pages 150–161, Cham, 2024. Springer Nature Switzerland. doi:10.1007/978-3-031-59167-9_13.
- [47] Albert Palomer, Pere Ridao, Dina Youakim, David Ribas, Josep Forest, and Yvan Petillot. 3D laser scanner for underwater manipulation. *Sensors (Switzerland)*, 18(4):1–18, 2018. doi:10.3390/s18041086.
- [48] Maria Inês Pereira and Andry Maykol Pinto. Reinforcement learning based robot navigation using illegal actions for autonomous docking of surface vehicles in unknown environments. *Engineering Applications of Artificial Intelligence*, 133:108506, 2024. doi:10.1016/j.engappai.2024.108506.
- [49] Rafael Marques Claro, Diogo Brandão Silva, and Andry Maykol Pinto. Artuga: A novel multimodal fiducial marker for aerial robotics. *Robotics and Autonomous Systems*, 163:104398, 2023. doi:10.1016/j.robot.2023.104398.
- [50] Francisco Soares Neves, Rafael Marques Claro, and Andry Maykol Pinto. End-to-end detection of a landing platform for offshore uavs based on a multimodal early fusion approach. *Sensors*, 23(5), 2023. doi:10.3390/s23052434.
- [51] Daniel F. Campos, Eduardo P. Gonçalves, Hugo J. Campos, Maria I. Pereira, and Andry M. Pinto. Nautilus: An autonomous surface vehicle with a multilayer software architecture for offshore inspection. *Journal of Field Robotics*, 41(4):966–990, 2024. doi:10.1002/rob.22304.
- [52] Daniel Filipe Campos, Aníbal Matos, and Andry Maykol Pinto. Multi-domain inspection of offshore wind farms using an autonomous surface vehicle. *SN Applied Sciences*, 3(455), 2021. doi:10.1007/s42452-021-04451-5.
- [53] Matteo Franchi, Alessandro Ridolfi, and Benedetto Allotta. Underwater navigation with 2d forward looking sonar: An adaptive unscented kalman filter-based strategy for auvs. *Journal of Field Robotics*, 38(3):355–385, 2021. doi:10.1002/rob.21991.
- [54] Zhaoxin Dong, Weisheng Yan, Rongxin Cui, Lei Lei, and Yaozhen He. An adaptive grouping sonar-inertial odometry for underwater navigation. *Ocean Engineering*, 294:116688, 2024. doi:10.1016/j.oceaneng.2024.116688.
- [55] Thomas Guerneve, Kartic Subr, and Yvan Petillot. Three-dimensional reconstruction of underwater objects using wide-aperture imaging sonar. *Journal of Field Robotics*, 35(6):890–905, 2018. doi:10.1002/rob.21783.
- [56] Pedro V. Teixeira, Dehann Fourie, Michael Kaess, and John J. Leonard. Dense, sonar-based reconstruction of underwater scenes. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8060–8066, 2019. doi:10.1109/IROS40897.2019.8968071.

- [57] Byeongjin Kim, Jason Kim, Hyeonwoo Cho, Jinwhan Kim, and Son-Cheol Yu. Auv-based multi-view scanning method for 3-d reconstruction of underwater object using forward scan sonar. *IEEE Sensors Journal*, 20(3):1592–1606, 2020. doi:10.1109/JSEN.2019.2946587.
- [58] Mohamad Qadri, Michael Kaess, and Ioannis Gkioulekas. Neural implicit surface reconstruction using imaging sonar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1040–1047, 2023. doi:10.1109/ICRA48891.2023.10161206.
- [59] Yunxuan Feng, Wenjie Lu, Haowen Gao, Binyu Nie, Kaiyang Lin, and Liang Hu. Differentiable space carving for 3d reconstruction using imaging sonar. *IEEE Robotics and Automation Letters*, 9(11):10065–10072, 2024. doi:10.1109/LRA.2024.3469778.
- [60] Wenhao Xu, Jianmin Yang, Handi Wei, Haining Lu, Xinliang Tian, and Xin Li. Seabed mapping for deep-sea mining vehicles based on forward-looking sonar. *Ocean Engineering*, 299:117276, 2024. doi:10.1016/j.oceaneng.2024.117276.
- [61] Francisco Bonin-Font, Gabriel Oliver, Stephan Wirth, Miquel Massot, Pep Lluís Negre, and Joan Pau Beltran. Visual sensing for autonomous underwater exploration and intervention tasks. *Ocean Engineering*, 93:25–44, 2015. doi:10.1016/j.oceaneng.2014.11.005.
- [62] Amin Sarafriz and Brian K. Haus. A structured light method for underwater surface reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:40–52, 2016. doi:10.1016/j.isprsjprs.2016.01.014.
- [63] Ayoun Kim and Ryan M. Eustice. Real-time visual slam for autonomous underwater hull inspection using visual saliency. *IEEE Transactions on Robotics*, 29(3):719–733, 2013. doi:10.1109/TRO.2012.2235699.
- [64] Petter Risholm, Jens T. Thielemann, Richard Moore, and Karl Henrik Haugholt. A scatter removal technique to enhance underwater range-gated 3d and intensity images. In *OCEANS 2018 MTS/IEEE Charleston*, pages 1–6, 2018. doi:10.1109/OCEANS.2018.8604613.
- [65] Marc Carreras, Arnau Carrera, Narcís Palomeras, David Ribas, Natàlia Hurtós, Quim Salvi, and Pere Ridao. Intervention payload for valve turning with an auv. In Roberto Moreno-Díaz, Franz Pichler, and Alexis Quesada-Arencibia, editors, *Computer Aided Systems Theory – EUROCAST 2015*, pages 877–884, Cham, 2015. Springer International Publishing. doi:10.1007/978-3-319-27340-2_107.
- [66] Narcís Palomeras, Antonio Peñalver, Miquel Massot-Campos, Guillem Vallicrosa, Pep Lluís Negre, J. Javier Fernández, Pere Ridao, Pedro J. Sanz, Gabriel Oliver-Codina, and Albert Palomer. I-auv docking and intervention in a subsea panel. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2279–2285, 2014. doi:10.1109/IROS.2014.6942870.
- [67] Mario Prats, David Ribas, Narcís Palomeras, Juan Carlos García, Volker Nannen, Stephan Wirth, J. Javier Fernández, Joan P. Beltrán, Ricard Campos, Pere Ridao, Pedro J. Sanz, Gabriel Oliver, Marc Carreras, Nuno Gracias, Raúl Marín, and Alberto Ortiz. Reconfigurable AUV for intervention missions: A case study on underwater object recovery. *Intelligent Service Robotics*, 5(1):19–31, 2012. doi:10.1007/s11370-011-0101-z.
- [68] Satja Sivčev, Matija Rossi, Joseph Coleman, Edin Omerdić, Gerard Dooly, and Daniel Toal. Collision detection for underwater rovs manipulator systems. *Sensors*, 18(4), 2018. doi:10.3390/s18041117.

- [69] Yan-Tsung Peng and Pamela C. Cosman. Underwater image restoration based on image blurriness and light absorption. *IEEE Transactions on Image Processing*, 26(4):1579–1594, 2017. doi:10.1109/TIP.2017.2663846.
- [70] Raymond C. Smith and Karen S. Baker. Optical properties of the clearest natural waters (200–800 nm). *Applied Optics*, 20(2):177–184, 1981. doi:10.1364/AO.20.000177.
- [71] Sundara Tejaswi Digumarti, Gaurav Chaurasia, Aparna Taneja, Roland Siegwart, Amber Thomas, and Paul Beardsley. Underwater 3d capture using a low-cost commercial depth camera. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. doi:10.1109/WACV.2016.7477644.
- [72] Yuanjie Zheng, Stephen Lin, Sing Bing Kang, Rui Xiao, James C. Gee, and Chandra Kambhamettu. Single-image vignetting correction from gradient distribution symmetries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1480–1494, 2013. doi:10.1109/TPAMI.2012.210.
- [73] Huimin Lu, Yujie Li, Xing Xu, Jianru Li, Zhifei Liu, Xin Li, Jianmin Yang, and Seiichi Serikawa. Underwater image enhancement method using weighted guided trigonometric filtering and artificial light correction. *Journal of Visual Communication and Image Representation*, 38:504–516, 2016. doi:10.1016/j.jvcir.2016.03.029.
- [74] Mengyao Li, Kun Wang, Liquan Shen, Yufei Lin, Zhengyong Wang, and Qijie Zhao. Uialn: Enhancement for underwater image with artificial light. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3622–3637, 2023. doi:10.1109/TCSVT.2023.3237993.
- [75] Mitch Bryson, Matthew Johnson-Roberson, Oscar Pizarro, and Stefan B. Williams. True color correction of autonomous underwater vehicle imagery. *Journal of Field Robotics*, 33(6):853–874, 2016. doi:10.1002/rob.21638.
- [76] Monika Roznere and Alberto Quattrini Li. Real-time model-based image color correction for underwater robots. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7191–7196, 2019. doi:10.1109/IROS40897.2019.8967557.
- [77] Changli Li, Shiqiang Tang, Hon Keung Kwan, Jingwen Yan, and Teng Zhou. Color correction based on cfa and enhancement based on retinex with dense pixels for underwater images. *IEEE Access*, 8:155732–155741, 2020. doi:10.1109/ACCESS.2020.3019354.
- [78] Zheng Liang, Yafei Wang, Xueyan Ding, Zetian Mi, and Xianping Fu. Single underwater image enhancement by attenuation map guided color correction and detail preserved dehazing. *Neurocomputing*, 425:160–172, 2021. doi:10.1016/j.neucom.2020.03.091.
- [79] Jie Li, Katherine A. Skinner, Ryan M. Eustice, and Matthew Johnson-Roberson. Wattergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics and Automation Letters*, 3(1):387–394, 2018. doi:10.1109/LRA.2017.2730363.
- [80] Tianyi Zhang and Matthew Johnson-Roberson. Beyond nerf underwater: Learning neural reflectance fields for true color correction of marine imagery. *IEEE Robotics and Automation Letters*, 8(10):6467–6474, 2023. doi:10.1109/LRA.2023.3307287.

- [81] Iuliu Vasilescu, Carrick Detweiler, and Daniela Rus. Color-accurate underwater imaging using perceptual adaptive illumination. *Autonomous Robots*, 31(2-3):285–296, 2011. doi:10.1007/s10514-011-9245-0.
- [82] Ying Shen, Chuanjiang Zhao, Yu Liu, Shu Wang, and Feng Huang. Underwater optical imaging: Key technologies and applications review. *IEEE Access*, 9:85500–85514, 2021. doi:10.1109/ACCESS.2021.3086820.
- [83] Raimondo Schettini and Silvia Corchs. Underwater image processing: State of the art of restoration and image enhancement methods. *EURASIP Journal on Advances in Signal Processing*, 2010(1), 2010. doi:10.1155/2010/746052.
- [84] B. L. McGlamery. A Computer Model For Underwater Camera Systems. In Seibert Quimby Duntley, editor, *Ocean Optics VI*, volume 0208, pages 221 – 231. International Society for Optics and Photonics, SPIE, 1980. doi:10.1117/12.958279.
- [85] J. S. Jaffe. Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering*, 15(2):101–111, 1990. doi:10.1109/48.50695.
- [86] Petter Risholm, Jostein Thorstensen, Jens T. Thielemann, Kristin Kaspersen, Jon Tschudi, Chris Yates, Chris Softley, Igor Abrosimov, Jonathan Alexander, and Karl Henrik Haugholt. Real-time super-resolved 3d in turbid water using a fast range-gated cmos camera. *Applied Optics*, 57(14):3927–3937, May 2018. doi:10.1364/AO.57.003927.
- [87] Andry Maykol Pinto and Anibal C. Matos. Maresye: A hybrid imaging system for underwater robotic applications. *Information Fusion*, 55:16–29, 2020. doi:10.1016/j.inffus.2019.07.014.
- [88] Shuaibao Chen, Peng Liu, Wei He, Dong Luo, Yuguang Tan, Liangpei Chen, Jue Wang, Qi Zhao, Guohua Jiao, and Wei Chen. Polarization-enhanced underwater laser range-gated imaging for subaquatic applications. *Sensors*, 24(20), 2024. doi:10.3390/s24206681.
- [89] S. Selva Nidhyandhan, R. Sindhuja, and R. Shantha Selva Kumari. Double stage gaussian filter for better underwater image enhancement. *Wireless Personal Communications*, 114(4):2909–2921, 2020. doi:10.1007/s11277-020-07509-6.
- [90] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011. doi:10.1109/TPAMI.2010.168.
- [91] Yan-Tsung Peng, Keming Cao, and Pamela C. Cosman. Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing*, 27(6):2856–2868, 2018. doi:10.1109/TIP.2018.2813092.
- [92] Jingchun Zhou, Tongyu Yang, Weishen Chu, and Weishi Zhang. Underwater image restoration via backscatter pixel prior and color compensation. *Engineering Applications of Artificial Intelligence*, 111:104785, 2022. doi:10.1016/j.engappai.2022.104785.
- [93] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2020. doi:10.1109/TIP.2019.2955241.

- [94] Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 32:3066–3079, 2023. doi:10.1109/TIP.2023.3276332.
- [95] Young-Hoo Kwon. Object plane deformation due to refraction in two-dimensional underwater motion analysis. *Journal of Applied Biomechanics*, 15(4):396–403, November 1999. doi:10.1123/jab.15.4.396.
- [96] Anne Sedlazeck and Reinhard Koch. Perspective and non-perspective camera models in underwater imaging – overview and error analysis. In Frank Dellaert, Jan-Michael Frahm, Marc Pollefeys, Laura Leal-Taixé, and Bodo Rosenhahn, editors, *Outdoor and Large-Scale Real-World Scene Analysis*, pages 212–242, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. doi:10.1007/978-3-642-34091-8_10.
- [97] T. Treibitz, Y. Schechner, C. Kunz, and H. Singh. Flat refractive geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):51–65, 2012. doi:10.1109/TPAMI.2011.105.
- [98] Miquel Massot-Campos and Gabriel Oliver-Codina. Optical sensors and methods for underwater 3D reconstruction. *Sensors (Switzerland)*, 15(12):31525–31557, 2015. doi:10.3390/s151229864.
- [99] C. Kunz and H. Singh. Hemispherical refraction and camera calibration in underwater vision. In *OCEANS 2008*, pages 1–7, 2008. doi:10.1109/OCEANS.2008.5151967.
- [100] Fabio Menna, Erica Nocerino, Salvatore Troisi, and Fabio Remondino. A photogrammetric approach to survey floating and semi-submerged objects. In Fabio Remondino, Mark R. Shortis, Jürgen Beyerer, and Fernando Puente León, editors, *Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection*, volume 8791, pages 117 – 131. International Society for Optics and Photonics, SPIE, 2013. doi:10.1117/12.2020464.
- [101] Junfeng Fan, Xuan Wang, Chao Zhou, Yaming Ou, Fengshui Jing, and Zengguang Hou. Development, calibration, and image processing of underwater structured light vision system: A survey. *IEEE Transactions on Instrumentation and Measurement*, 72:1–18, 2023. doi:10.1109/TIM.2023.3235420.
- [102] Ricardo Ferreira, João P. Costeira, and João A. Santos. Stereo reconstruction of a submerged scene. In Jorge S. Marques, Nicolás Pérez de la Blanca, and Pedro Pina, editors, *Pattern Recognition and Image Analysis*, pages 102–109, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. doi:10.1007/11492429_13.
- [103] Lai Kang, Lingda Wu, and Yee-Hong Yang. Experimental study of the influence of refraction on underwater three-dimensional reconstruction using the svp camera model. *Appl. Opt.*, 51(31):7591–7603, 2012. doi:10.1364/AO.51.007591.
- [104] François Chadebecq, Francisco Vasconcelos, René Lacher, Efthymios Maneas, Adrien Desjardins, Sébastien Ourselin, Tom Vercauteren, and Danail Stoyanov. Refractive two-view reconstruction for underwater 3d vision. *International Journal of Computer Vision*, 128(5):1101–1117, November 2019. doi:10.1007/s11263-019-01218-9.
- [105] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *2020*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4022–4031, 2020. doi:10.1109/CVPR42600.2020.00408.
- [106] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12727–12736, 2021. doi:10.1109/ICCV48922.2021.01251.
- [107] R. Li, C. Tao, W. Zou, R.G. Smith, and T.A. Curran. An underwater digital photogrammetric system for fishery geomatics. *International Archives of Photogrammetry and Remote Sensing*, 31:319–323, 1996.
- [108] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. doi:10.1017/CBO9780511811685.
- [109] M. R. Shortis and E. S. Harvey. Design and calibration of an underwater stereo-video system for the monitoring of marine fauna populations. *International Archives of Photogrammetry and Remote Sensing*, 32(5):792–799, 1998.
- [110] Fabio Menna, Erica Nocerino, Salvatore Troisi, and Fabio Remondino. A photogrammetric approach to survey floating and semi-submerged objects. In Fabio Remondino, Mark R. Shortis, Jürgen Beyerer, and Fernando Puente León, editors, *Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection*, volume 8791, page 87910H. International Society for Optics and Photonics, SPIE, 2013. doi:10.1117/12.2020464.
- [111] Mark Shortis. Calibration techniques for accurate measurements by underwater camera systems. *Sensors (Switzerland)*, 15(12):30810–30827, 2015. doi:10.3390/s151229831.
- [112] Anne Jordt and Reinhard Koch. Calibration of housing parameters for underwater stereo-camera rigs. In *British Machine Vision Conference*, 2011. doi:10.5244/C.25.118.
- [113] C. Bräuer-Burchardt, C. Munkelt, M. Heinze, I. Gebhart, P. Kühmstedt, and G. Notni. Underwater 3d measurements with advanced camera modelling. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 90(1):55–67, 2022. doi:10.1007/s41064-022-00195-y.
- [114] 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 (CVPR)*, pages 3346–3353, 2012. doi:10.1109/CVPR.2012.6248073.
- [115] Xida Chen and Yee-Hong Yang. Two-view camera housing parameters calibration for multi-layer flat refractive interface. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 524–531, 2014. doi:10.1109/CVPR.2014.74.
- [116] Mengkun She, David Nakath, Yifan Song, and Kevin Köser. Refractive geometry for underwater domes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:525–540, 2022. doi:10.1016/j.isprsjprs.2021.11.006.
- [117] Timothy Yau, Minglun Gong, and Yee-Hong Yang. Underwater camera calibration using wavelength triangulation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2499–2506, 2013. doi:10.1109/CVPR.2013.323.
- [118] Changjun Gu, Yang Cong, Gan Sun, Yajun Gao, Xu Tang, Tao Zhang, and Baojie Fan. Meducc: Medium-driven underwater camera calibration for refractive 3-d reconstruction.

- IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(9):5937–5948, 2022. doi:10.1109/TSMC.2021.3132146.
- [119] Monika Roznere, Adithya K. Pediredla, Samuel E. Lensgraf, Yogesh Girdhar, and Alberto Quattrini Li. Underwater dome-port camera calibration: Modeling of refraction and offset through n-sphere camera model. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6110–6117, 2024. doi:10.1109/ICRA57147.2024.10611338.
- [120] Gianfranco Bianco, Alessandro Gallo, Fabio Bruno, and Maurizio Muzzupappa. A comparative analysis between active and passive techniques for underwater 3D reconstruction of close-range objects. *Sensors (Switzerland)*, 13(8):11007–11031, 2013. doi:10.3390/s130811007.
- [121] David Eigen, Christian Puhersch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14*, page 2366–2374, Cambridge, MA, USA, 2014. MIT Press. doi:10.5555/2969033.2969091.
- [122] Ali Karami, Fabio Menna, and Fabio Remondino. Combining photogrammetry and photometric stereo to achieve precise and complete 3d reconstruction. *Sensors*, 22(21), 2022. doi:10.3390/s22218172.
- [123] A. G. Allais, V. Brandou, Ulrich Hoge, Melanie Bergmann, J. P. L  v  que, P. L  on, J. F. Cadiou, J. Sarrazin, and P. M. Sarradin. Design of optical instrumentation for 3d and temporal deep-sea observation. In *MarTech workshop, Bremerhaven/Bremen, Germany*, 2005.
- [124] V. Brandou, A. G. Allais, M. Perrier, E. Malis, P. Rives, J. Sarrazin, and P. M. Sarradin. 3d reconstruction of natural underwater scenes using the stereovision system iris. In *OCEANS 2007 - Europe*, pages 1–6, 2007. doi:10.1109/OCEANSE.2007.4302315.
- [125] Andrew Hogue, Andrew German, and Michael Jenkin. Underwater environment reconstruction using stereo and inertial data. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, pages 2372–2377, 2007. doi:10.1109/ICSMC.2007.4413666.
- [126] Pep Luis Negre Carrasco, Francisco Bonin-Font, Miquel Massot Campos, and Gabriel Oliver Codina. Stereo-vision graph-SLAM for robust navigation of the AUV SPARUS II. *IFAC-PapersOnLine*, 28(2):200–205, 2015. doi:10.1016/j.ifacol.2015.06.033.
- [127] Kresimir Williams, Alex De Robertis, Zachary Berkowitz, Chris Rooper, and Rick Towler. An underwater stereo-camera trap. *Methods in Oceanography*, 11(2014):1–12, 2014. doi:10.1016/j.mio.2015.01.003.
- [128] Fabio Oleari, Fabjan Kallasi, Dario Lodi Rizzini, Jacopo Aleotti, and Stefano Caselli. An underwater stereo vision system: From design to deployment and dataset acquisition. *MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World*, 2015. doi:10.1109/OCEANS-Genova.2015.7271529.
- [129] Matthew Johnson-Roberson, Mitch Bryson, Ariell Friedman, Oscar Pizarro, Giancarlo Troni, Paul Ozog, and Jon C. Henderson. High-resolution underwater robotic vision-based mapping and three-dimensional reconstruction for archaeology. *Journal of Field Robotics*, 34(4):625–643, 2017. doi:10.1002/rob.21658.

- [130] Guanying Huo, Ziyin Wu, Jiabiao Li, and Shoujun Li. Underwater target detection and 3d reconstruction system based on binocular vision. *Sensors*, 18(10), 2018. doi:10.3390/s18103570.
- [131] Myo Myint, Kenta Yonemori, Khin Nwe Lwin, Akira Yanou, and Mamoru Minami. Dual-eyes vision-based docking system for autonomous underwater vehicle: an approach and experiments. *Journal of Intelligent and Robotic Systems: Theory and Applications*, 92:159–186, 9 2018. doi:10.1007/s10846-017-0703-6.
- [132] Emma V. Sheehan, Danielle Bridger, Sarah J. Nancollas, and Simon J. Pittman. PelagiCam: a novel underwater imaging system with computer vision for semi-automated monitoring of mobile marine fauna at offshore structures. *Environmental Monitoring and Assessment*, 192(1), 2020. doi:10.1007/s10661-019-7980-4.
- [133] Yunpeng Ma, Yi Wu, Qingwu Li, Yaqin Zhou, and Dabing Yu. Rov-based binocular vision system for underwater structure crack detection and width measurement. *Multimedia Tools and Applications*, 82:20899–20923, 6 2023. doi:10.1007/s11042-022-14168-1.
- [134] Weihan Wang, Bharat Joshi, Nathaniel Burgdorfer, Konstantinos Batsosc, Alberto Quattrini Lid, Philippos Mordohaia, and Ioannis Rekleitisb. Real-time dense 3d mapping of underwater environments. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5184–5191, 2023. doi:10.1109/ICRA48891.2023.10160266.
- [135] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. doi:10.1109/CVPR.2016.445.
- [136] Simone Tani, Francesco Ruscio, Matteo Bresciani, Bo Miquel Nordfeldt, Francisco Bonin-Font, and Riccardo Costanzi. Development and testing of a navigation solution for autonomous underwater vehicles based on stereo vision. *Ocean Engineering*, 280:114757, 2023. doi:10.1016/j.oceaneng.2023.114757.
- [137] Chen Yuehang, Lin Dongyun, Lan WeiYao, and Liu Tengfei. Underwater measurement of abalone shell’s size based on stereo vision. *Measurement*, 242:115699, 2025. doi:10.1016/j.measurement.2024.115699.
- [138] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5410–5418. IEEE, June 2018. doi:10.1109/cvpr.2018.00567.
- [139] Eric Alfaro-Dufour, Caterina Muntaner-González, Antoni Martorell-Torres, and Gabriel Oliver-Codina. Lanty: A deep sea stereo vision system. In *2024 IEEE International Conference on Industrial Technology (ICIT)*, pages 1–6, 2024. doi:10.1109/ICIT58233.2024.10540725.
- [140] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. doi:10.1109/TPAMI.2007.1166.
- [141] César Silva, Ricardo Aires, and Flávio Rodrigues. A compact underwater stereo vision system for measuring fish. *Aquaculture and Fisheries*, 9(6):1000–1006, 2024. doi:10.1016/j.aaf.2023.03.006.

- [142] Kai Hu, Tianyan Wang, Chaowen Shen, Chenghang Weng, Fenghua Zhou, Min Xia, and Liguu Weng. Overview of underwater 3d reconstruction technology based on optical images. *Journal of Marine Science and Engineering*, 11(5), 2023. doi:10.3390/jmse11050949.
- [143] Kemal E. Ozden, Konrad Schindler, and Luc Van Gool. Multibody structure-from-motion in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1134–1141, 2010. doi:10.1109/TPAMI.2010.23.
- [144] Simone Bianco, Gianluigi Ciocca, and Davide Marelli. Evaluating the performance of structure from motion pipelines. *Journal of Imaging*, 4(8):98, 2018. doi:10.3390/jimaging4080098.
- [145] Anne Sedlazeck, Kevin Köser, and Reinhard Koch. 3D reconstruction based on underwater video from ROV Kiel 6000 considering underwater imaging conditions. *OCEANS '09 IEEE Bremen: Balancing Technology with Future Needs*, 2009. doi:10.1109/OCEANSE.2009.5278305.
- [146] Chris Beall, Brian J. Lawrence, Viorela Ila, and Frank Dellaert. 3d reconstruction of underwater structures. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4418–4423, 2010. doi:10.1109/IROS.2010.5649213.
- [147] David McKinnon, Hu He, Ben Upcroft, and Ryan N. Smith. Towards automated and in-situ, near-real time 3-D reconstruction of coral reef environments. *OCEANS'11 - MTS/IEEE Kona, Program Book*, 2011. doi:10.23919/oceans.2011.6106982.
- [148] Anne Jordt-Sedlazeck and Reinhard Koch. Refractive structure-from-motion on underwater images. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 57–64, 2013. doi:10.1109/ICCV.2013.14.
- [149] Akira Shibata, Hiromitsu Fujii, Atsushi Yamashita, and Hajime Asama. Scale-reconstructable structure from motion using a refractive plate and a single camera. *Seimitsu Kogaku Kaishi/Journal of the Japan Society for Precision Engineering*, 82(12):1045–1053, 2016. doi:10.2493/jjspe.82.1045.
- [150] Xiao Xu, Renzheng Che, Rui Nian, Bo He, Meimei Chen, and Amaury Lendasse. Underwater 3d object reconstruction with multiple views in video stream via structure from motion. In *OCEANS 2016 - Shanghai*, pages 1–5, 2016. doi:10.1109/OCEANSAP.2016.7485694.
- [151] François Chadebecq, Francisco Vasconcelos, George Dwyer, René Lacher, Sébastien Ourselin, Tom Vercauteren, and Danail Stoyanov. Refractive structure-from-motion through a flat refractive interface. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5325–5333, 2017. doi:10.1109/ICCV.2017.568.
- [152] François Chadebecq, Francisco Vasconcelos, René Lacher, Efthymios Maneas, Adrien Desjardins, Sébastien Ourselin, Tom Vercauteren, and Danail Stoyanov. Refractive Two-View Reconstruction for Underwater 3D Vision. *International Journal of Computer Vision*, 128(5):1101–1117, 2020. doi:10.1007/s11263-019-01218-9.
- [153] Klemen Istenič, Nuno Gracias, Aurélien Arnaubec, Javier Escartín, and Rafael Garcia. Automatic scale estimation of structure from motion based 3D models using laser scalers in underwater scenarios. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159(November 2019):13–25, 2020. doi:10.1016/j.isprsjprs.2019.10.007.

- [154] Yanli Chen, Qiushi Li, Shenghua Gong, Jun Liu, and Wenxue Guan. Uv3d: Underwater video stream 3d reconstruction based on efficient global sfm. *Applied Sciences*, 12(12):5918, 2022. doi:10.3390/app12125918.
- [155] Mengkun She, Felix Seegräber, David Nakath, and Kevin Köser. Refractive colmap: Refractive structure-from-motion revisited, 2024. doi:10.48550/ARXIV.2403.08640.
- [156] Hayato Kondo, Toshihiro Maki, Tamaki Ura, Yoshiaki Nose, Takashi Sakamaki, and Masaaki Inaishi. Structure tracing with a ranging system using a sheet laser beam. *2004 International Symposium on Underwater Technology, UT'04 - Proceedings*, pages 83–88, 2004. doi:10.1109/ut.2004.1405483.
- [157] A. Yamashita, H. Higuchi, T. Kaneko, and Y. Kawata. Three dimensional measurement of object's surface in water using the light stripe projection method. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 3, pages 2736–2741. IEEE, 2004. doi:10.1109/robot.2004.1307474.
- [158] Gabrielle Inglis, Clara Smart, Ian Vaughn, and Chris Roman. A pipeline for structured light bathymetric mapping. *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 4425–4432, 2012. doi:10.1109/IROS.2012.6386038.
- [159] Adrian Bodenmann, Blair Thornton, Ryota Nakajima, Hiroyuki Yamamoto, and Tamaki Ura. Wide area 3d seafloor reconstruction and its application to sea fauna density mapping. In *2013 OCEANS - San Diego*, pages 1–5, 2013. doi:10.23919/OCEANS.2013.6741280.
- [160] Miquel Massot-Campos and Gabriel Oliver-Codina. Underwater laser-based structured light system for one-shot 3d reconstruction. In *IEEE SENSORS 2014 Proceedings*. IEEE, 2014. doi:10.1109/icsens.2014.6985208.
- [161] Nicholas Hansen, Mikkel C. Nielsen, David Johan Christensen, and Mogens Blanke. Short-range sensor for underwater robot navigation using line-lasers and vision. *IFAC-PapersOnLine*, 28(16):113–120, 2015. doi:10.1016/j.ifacol.2015.10.267.
- [162] Christos C. Constantinou, Savvas G. Loizou, and George P. Georgiades. An underwater laser vision system for relative 3-d posture estimation to mesh-like targets. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2036–2041, 2016. doi:10.1109/IROS.2016.7759320.
- [163] M. Bleier and A. Nüchter. Low-cost 3d laser scanning in air or water using self-calibrating structured light. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W3:105–112, 2017. doi:10.5194/isprs-archives-XLII-2-W3-105-2017.
- [164] Gabriel Matos, Pedro Buschinelli, and Tiago Pinto. Underwater laser triangulation sensor model with flat refractive interfaces. *IEEE Journal of Oceanic Engineering*, 45(3):937–945, 2020. doi:10.1109/JOE.2019.2891863.
- [165] Shuaishuai Li, Xiang Gao, and Zexiao Xie. Underwater structured light stripe center extraction with normalized grayscale gravity method. *Sensors*, 23(24):9839, 2023. doi:10.3390/s23249839.
- [166] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017. doi:10.1109/CVPR.2017.660.

- [167] Junfeng Fan, Yaming Ou, Xuan Li, Chao Zhou, and Zengguang Hou. Structured light vision based pipeline tracking and 3d reconstruction method for underwater vehicle. *IEEE Transactions on Intelligent Vehicles*, 9(2):3372–3383, 2024. doi:10.1109/TIV.2023.3340737.
- [168] Shuaishuai Li, Xiang Gao, and Zexiao Xie. Binocular underwater measurement with multicolor structured light. *IEEE Journal of Oceanic Engineering*, 49(2):649–666, 2024. doi:10.1109/JOE.2023.3315397.
- [169] Joaquim Salvi, Jordi Pagès, and Joan Batlle. Pattern codification strategies in structured light systems. *Pattern Recognition*, 37(4):827–849, 2004. doi:10.1016/j.patcog.2003.10.002.
- [170] F. Bruno, G. Bianco, M. Muzzupappa, S. Barone, and A. V. Razionale. Experimentation of structured light and stereo vision for underwater 3D reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(4):508–518, 2011. doi:10.1016/j.isprsjprs.2011.02.009.
- [171] Sufeng Zhuang, Dawei Tu, Xu Zhang, and Chuzhuang Liu. The influence of active projection speckle patterns on underwater binocular stereo vision 3d imaging. *Optics Communications*, 528:129014, 2023. doi:10.1016/j.optcom.2022.129014.
- [172] Qican Zhang, Qingfeng Wang, Zhiling Hou, Yuankun Liu, and Xianyu Su. Three-dimensional shape measurement for an underwater object based on two-dimensional grating pattern projection. *Optics and Laser Technology*, 43(4):801–805, 2011. doi:10.1016/j.optlastec.2010.11.007.
- [173] Hiroki Morinaga, Hirohisa Baba, Marco Visentini-Scarzanella, Hiroshi Kawasaki, Ryo Furukawa, and Ryusuke Sagawa. Underwater active oneshot scan with static wave pattern and bundle adjustment. In *Image and Video Technology: 7th Pacific-Rim Symposium*, page 404–418, Berlin, Heidelberg, 2015. Springer-Verlag. doi:10.1007/978-3-319-29451-3_33.
- [174] Y. Nishida, S. Yasukawa, and K. Ishii. Underwater 3d scanner using rgb laser pattern. In *2021 IEEE/SICE International Symposium on System Integration (SII)*, pages 229–232, 2021. doi:10.1109/IEEECONF49454.2021.9382643.
- [175] Zhiliang Xu, Lizhuang Ma, and Wuzheng Tan. Robust dense depth acquisition using 2-d de bruijn structured light. In Lizhuang Ma, Matthias Rauterberg, and Ryohei Nakatsu, editors, *Entertainment Computing – ICEC 2007*, pages 304–314, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. doi:10.1007/978-3-540-74873-1_37.
- [176] Sufeng Zhuang, Dawei Tu, Xu Zhang, and Chuzhuang Liu. The influence of active projection speckle patterns on underwater binocular stereo vision 3d imaging. *Optics Communications*, 528:129014, 2023. doi:10.1016/j.optcom.2022.129014.
- [177] Flavio Lopes, Hugo Silva, Jose Miguel Almeida, Alfredo Martins, and Eduardo Silva. Structured light system for underwater inspection operations. *MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World*, pages 1–6, 2015. doi:10.1109/OCEANS-Genova.2015.7271564.
- [178] Shukai Chi, Zexiao Xie, and Wenzhu Chen. A laser line auto-scanning system for underwater 3d reconstruction. *Sensors*, 16(9):1534, 2016. doi:10.3390/s16091534.

- [179] Miguel Castellón, Albert Palomer, Josep Forest, and Pere Ridao. Underwater 3d scanner model using a biaxial mems mirror. *IEEE Access*, 9:50231–50243, 2021. doi:10.1109/ACCESS.2021.3069189.
- [180] Qingsheng Xue, Qian Sun, Fupeng Wang, Haoxuan Bai, Bai Yang, and Qian Li. Underwater high-precision 3d reconstruction system based on rotating scanning. *Sensors*, 21(4):1402, 2021. doi:10.3390/s21041402.
- [181] Changjun Gu, Yang Cong, and Gan Sun. Three birds, one stone: Unified laser-based 3-d reconstruction across different media. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021. doi:10.1109/TIM.2020.3039641.
- [182] Yaming Ou, Junfeng Fan, Chao Zhou, Shifei Tian, Long Cheng, and Min Tan. Binocular structured light 3-d reconstruction system for low-light underwater environments: Design, modeling, and laser-based calibration. *IEEE Transactions on Instrumentation and Measurement*, 72:1–14, 2023. doi:10.1109/TIM.2023.3261941.
- [183] Shuaishuai Li, Xiang Gao, Haoyue Wang, and Zexiao Xie. Monocular underwater measurement of structured light by scanning with vibrating mirrors. *Optics and Lasers in Engineering*, 169:107738, 2023. doi:10.1016/j.optlaseng.2023.107738.
- [184] Haitao Lin, Hua Zhang, Yonglong Li, Jianwen Huo, Hao Deng, and Huan Zhang. Method of 3d reconstruction of underwater concrete by laser line scanning. *Optics and Lasers in Engineering*, 183:108468, 2024. doi:10.1016/j.optlaseng.2024.108468.
- [185] Haitao Lin, Yonglong Li, Hua Zhang, Jianwen Huo, Jialong Li, and Huan Zhang. Integration of line structured light and stereo vision for underwater concrete 3d reconstruction. *SSRN*, 2024. doi:10.2139/ssrn.4751399.
- [186] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4009–4018, June 2021. doi:10.1109/CVPR46437.2021.00400.
- [187] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. doi:10.48550/arXiv.2302.12288.
- [188] P. L. J. Drews, E. R. Nascimento, S. S. C. Botelho, and M. F. Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE Computer Graphics and Applications*, 36(2):24–35, 2016. doi:10.1109/MCG.2016.26.
- [189] Junting Wang, Xiufen Ye, Yusong Liu, Xinkui Mei, and Jun Hou. Underwater self-supervised monocular depth estimation and its application in image enhancement. *Engineering Applications of Artificial Intelligence*, 120:105846, 2023. doi:10.1016/j.engappai.2023.105846.
- [190] H. Gupta and K. Mitra. Unsupervised single image underwater depth estimation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 624–628, 2019. doi:10.1109/ICIP.2019.8804200.

- [191] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2822–2837, 2021. doi:10.1109/TPAMI.2020.2977624.
- [192] Xinchun Ye, Zheng Li, Baoli Sun, Zhihui Wang, Rui Xu, Haojie Li, and Xin Fan. Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3995–4008, 2020. doi:10.1109/TCSVT.2019.2958950.
- [193] Javier Pérez, Mitch Bryson, Stefan B. Williams, and Pedro J. Sanz. Recovering depth from still images for underwater dehazing using deep learning. *Sensors*, 20(16):4580, 2020. doi:10.3390/s20164580.
- [194] Praful Hambarde, Subrahmanyam Murala, and Abhinav Dhall. Uw-gan: Single-image depth estimation and image enhancement for underwater images. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021. doi:10.1109/TIM.2021.3120130.
- [195] Xuewen Yang, Xing Zhang, Nan Wang, Guoling Xin, and Wenjie Hu. Underwater self-supervised depth estimation. *Neurocomputing*, 514:362–373, 2022. doi:10.1016/j.neucom.2022.09.122.
- [196] 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, 2019. doi:10.1109/CVPR.2019.00178.
- [197] Boxiao Yu, Jiayi Wu, and Md Jahidul Islam. Udepth: Fast monocular depth estimation for visually-guided underwater robots. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3116–3123, 2023. doi:10.1109/ICRA48891.2023.10161471.
- [198] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. doi:10.1109/CVPR.2018.00474.
- [199] Lin Hong, Xin Wang, Gan Zhang, and Ming Zhao. Usod10k: A new benchmark dataset for underwater salient object detection. *IEEE Transactions on Image Processing*, pages 1–1, 2023. doi:10.1109/TIP.2023.3266163.
- [200] Chen Wang, Haiyong Xu, Gangyi Jiang, Mei Yu, Ting Luo, and Yeyao Chen. Underwater monocular depth estimation based on physical-guided transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi:10.1109/TGRS.2024.3373904.
- [201] Jian Wang, Jing Wang, Shenghui Rong, and Bo He. Umono: Physical model informed hybrid cnn-transformer framework for underwater monocular depth estimation, 2024. doi:10.48550/arXiv.2407.17838.
- [202] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with stable diffusion, 2023. doi:10.48550/arXiv.2312.12471.
- [203] Alzayat Saleh, Melanie Olsen, Bouchra Senadji, and Mostafa Rahimi Azghadi. Learning from the giants: A practical approach to underwater depth and surface normals estimation, 2024. doi:10.48550/arXiv.2410.02072.

- [204] Yilin Ding, Kunqian Li, Han Mei, Shuaixin Liu, and Guojia Hou. Watermono: Teacher-guided anomaly masking and enhancement boosting for robust underwater self-supervised monocular depth estimation, 2024. doi:10.48550/arXiv.2406.13344.
- [205] Yelena Randall and Tali Treibitz. Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets, 2023. doi:10.48550/arXiv.2302.12772.
- [206] Kazuto Ichimaru, Ryo Furukawa, and Hiroshi Kawasaki. Multi-scale CNN stereo and pattern removal technique for underwater active stereo system. *Proceedings - 2018 International Conference on 3D Vision (3DV)*, pages 61–70, 2018. doi:10.1109/3DV.2018.00018.
- [207] Katherine A. Skinner, Junming Zhang, Elizabeth A. Olson, and Matthew Johnson-Roberson. UW StereoNet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery. *Proceedings - IEEE International Conference on Robotics and Automation (ICRA)*, pages 7947–7954, 2019. doi:10.1109/ICRA.2019.8794272.
- [208] Xinchun Ye, Jinyi Zhang, Yazhi Yuan, Rui Xu, Zhihui Wang, and Haojie Li. Underwater depth estimation via stereo adaptation networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5089–5101, 2023. doi:10.1109/TCSVT.2023.3249223.
- [209] Qingxuan Lv, Junyu Dong, Yuezun Li, Sheng Chen, Hui Yu, Shu Zhang, and Wenhan Wang. Uwstereo: A large synthetic dataset for underwater stereo matching, 2024. doi:10.48550/arXiv.2409.01782.
- [210] Qing Li, Hongjian Wang, Yao Xiao, Hualong Yang, Zhikang Chi, and Dongchen Dai. Underwater unsupervised stereo matching method based on semantic attention. *Journal of Marine Science and Engineering*, 12(7):1123, 2024. doi:10.3390/jmse12071123.
- [211] Will Maddern and Paul Newman. Real-time probabilistic fusion of sparse 3D LIDAR and dense stereo. *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 2181–2188, 2016. doi:10.1109/IROS.2016.7759342.
- [212] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. doi:10.1109/CVPR.2012.6248074.
- [213] Muhammad Kashif Ali, Asif Rajput, Muhammad Shahzad, Farhan Khan, Faheem Akhtar, and Anko Borner. Multi-sensor depth fusion framework for real-time 3d reconstruction. *IEEE Access*, 7:136471–136480, 2019. doi:10.1109/ACCESS.2019.2942375.
- [214] Yufeng Yue, Chule Yang, Yuanzhe Wang, P. G. Chaminda Namal Senarathne, Jun Zhang, Mingxing Wen, and Danwei Wang. A Multilevel Fusion System for Multirobot 3-D Mapping Using Heterogeneous Sensors. *IEEE Systems Journal*, 14(1):1341–1352, 2020. doi:10.1109/JSYST.2019.2927042.
- [215] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. doi:10.1109/MSP.2017.2738401.
- [216] Kihong Park, Seungryong Kim, and Kwanghoon Sohn. High-precision depth estimation with the 3d lidar and stereo fusion. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2156–2163, 2018. doi:10.1109/ICRA.2018.8461048.

- [217] Xuelian Cheng, Yiran Zhong, Yuchao Dai, Pan Ji, and Hongdong Li. Noise-aware unsupervised deep lidar-stereo fusion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6332–6341, 2019. doi:10.1109/CVPR.2019.00650.
- [218] Diogo Martins, Kevin Van Hecke, and Guido De Croon. Fusion of stereo and still monocular depth estimates in a self-supervised learning context. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 849–856, 2018. doi:10.1109/ICRA.2018.8461116.
- [219] Zuria Bauer, Felix Escalona, Edmanuel Cruz, Miguel Cazorla, and Francisco Gomez-Donoso. Refining the fusion of pepper robot and estimated depth maps method for improved 3d perception. *IEEE Access*, 7:185076–185085, 2019. doi:10.1109/ACCESS.2019.2960798.
- [220] Jaesung Choe, Kyungdon Joo, Tooba Imtiaz, and In So Kweon. Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation. *IEEE Robotics and Automation Letters*, 6(3):4672–4679, 2021. doi:10.1109/LRA.2021.3068712.
- [221] Haitao Meng, Changcai Li, Chonghao Zhong, Jianfeng Gu, Gang Chen, and Alois Knoll. Fastfusion: Deep stereo-lidar fusion for real-time high-precision dense depth sensing. *Journal of Field Robotics*, 40(7):1804–1816, 2023. doi:10.1002/rob.22216.
- [222] Jiayuan Fan, Haixiang Chen, Weide Liu, Xun Xu, and Jun Cheng. Holistic and contextual evidential stereo-lidar fusion for depth estimation. *IEEE Transactions on Intelligent Vehicles (TIV)*, pages 1–12, 2024. doi:10.1109/TIV.2024.3398210.
- [223] Jianxin Shi, Yuehua Li, Ziheng Zhang, Tiejun Li, and Jingbo Zhou. Adaptive weighted data fusion for line structured light and photometric stereo measurement system. *Sensors*, 24(13):4187, 2024. doi:10.3390/s24134187.
- [224] Jason Kim, Meungsuk Lee, Seokyong Song, Byeongjin Kim, and Son-Cheol Yu. 3-d reconstruction of underwater objects using image sequences from optical camera and imaging sonar. In *OCEANS 2019 MTS/IEEE SEATTLE*, pages 1–6, 2019. doi:10.23919/OCEANS40490.2019.8962558.
- [225] John McConnell, John D. Martin, and Brendan Englot. Fusing concurrent orthogonal wide-aperture sonar images for dense underwater 3d reconstruction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1653–1660, 2020. doi:10.1109/IROS45743.2020.9340995.
- [226] Shahriar Negahdaripour. Application of forward-scan sonar stereo for 3-d scene reconstruction. *IEEE Journal of Oceanic Engineering*, 45(2):547–562, 2020. doi:10.1109/JOE.2018.2875574.
- [227] Nicholas Sadjoli, Yiyu Cai, Gerald Seet, and Basman Elhadidi. Pcd reconstruction, object classification and pose estimation for underwater vehicles using orthogonal multibeam forward looking sonar fusion. *Ocean Engineering*, 288:116019, 2023. doi:10.1016/j.oceaneng.2023.116019.
- [228] Hangil Joe, Hyeonwoo Cho, Minsung Sung, Jinwhan Kim, and Son cheol Yu. Sensor fusion of two sonar devices for underwater 3d mapping with an AUV. *Autonomous Robots*, June 2021. doi:10.1007/s10514-021-09986-5.

- [229] Maria Cecilia Luna Alvarado, Fernando Darío Almeida García, Lenin Patricio Jiménez Jiménez, Gustavo Fraidenraich, and Yuzo Iano. Performance evaluation of soca-cfar detectors in weibull-distributed clutter environments. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi:10.1109/LGRS.2022.3152936.
- [230] Alexandre Cardaillac and Martin Ludvigsen. Camera-sonar combination for improved underwater localization and mapping. *IEEE Access*, 11:123070–123079, 2023. doi:10.1109/ACCESS.2023.3329834.
- [231] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. doi:10.1109/TRO.2015.2463671.
- [232] Jiawei Zhang, Fenglei Han, Duanfeng Han, Jianfeng Yang, Wangyuan Zhao, and Hansheng Li. Advanced underwater measurement system for rovs: Integrating sonar and stereo vision for enhanced subsea infrastructure maintenance. *Journal of Marine Science and Engineering*, 12(2):306, 2024. doi:10.3390/jmse12020306.
- [233] Cesar Cadena, Anthony Dick, and Ian D. Reid. Multi-modal auto-encoders as joint estimators for robotics scene understanding. *Robotics: Science and Systems*, 2016. doi:10.15607/rss.2016.xii.041.
- [234] Michele Mancini, Gabriele Costante, Paolo Valigi, and Thomas A. Ciarfuglia. Fast robust monocular depth estimation for Obstacle Detection with fully convolutional networks. *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 4296–4303, 2016. doi:10.1109/IROS.2016.7759632.
- [235] Yiyi Liao, Lichao Huang, Yue Wang, Sarath Kodagoda, Yinan Yu, and Yong Liu. Parse geometry from a line: Monocular depth estimation with partial laser observation. *Proceedings - IEEE International Conference on Robotics and Automation (ICRA)*, pages 5059–5066, 2017. doi:10.1109/ICRA.2017.7989590.
- [236] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: Depth and motion network for learning monocular stereo. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5622–5631, 2017. doi:10.1109/CVPR.2017.596.
- [237] Fangchang Mal and Sertac Karaman. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image. *Proceedings - IEEE International Conference on Robotics and Automation (ICRA)*, pages 4796–4803, 2018. doi:10.1109/ICRA.2018.8460184.
- [238] A. Hirata, R. Ishikawa, M. Roxas, and T. Oishi. Real-time dense depth estimation using semantically-guided lidar data propagation and motion stereo. *IEEE Robotics and Automation Letters*, 4(4):3806–3811, 2019. doi:10.1109/LRA.2019.2927126.
- [239] K. Wang and S. Shen. Flow-motion and depth network for monocular stereo and beyond. *IEEE Robotics and Automation Letters*, 5(2):3307–3314, 2020. doi:10.1109/LRA.2020.2975750.
- [240] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks

- for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021. doi:10.1109/LRA.2021.3060707.
- [241] Md Fahim Faysal Khan, Anusha Devulapally, Siddharth Advani, and Vijaykrishnan Narayanan. Robust multimodal depth estimation using transformer based generative adversarial networks. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3559–3568, New York, USA, 2022. Association for Computing Machinery. doi:10.1145/3503161.3548418.
- [242] Tomoyasu Shimada, Hiroki Nishikawa, Xiangbo Kong, and Hiroyuki Tomiyama. Fast and high-quality monocular depth estimation with optical flow for autonomous drones. *Drones*, 7(2):134, 2023. doi:10.3390/drones7020134.
- [243] Ji Zhang, Yiran Ding, and Zixin Liu. Occfusion: Depth estimation free multi-sensor fusion for 3d occupancy prediction, 2024. doi:10.48550/arXiv.2403.05329.
- [244] Baihui Xiao, Jingzehua Xu, Zekai Zhang, Tianyu Xing, Jingjing Wang, and Yong Ren. Multimodal monocular dense depth estimation with event-frame fusion using transformer. In Michael Wand, Kristína Malinová, Jürgen Schmidhuber, and Igor V. Tetko, editors, *Artificial Neural Networks and Machine Learning – ICANN 2024*, pages 419–433, Cham, 2024. Springer Nature Switzerland. doi:10.1007/978-3-031-72335-3_29.
- [245] Han Li, Yukai Ma, Yaqing Gu, Kewei Hu, Yong Liu, and Xingxing Zuo. Radarcam-depth: Radar-camera fusion for depth estimation with learned metric scale, 2024. doi:10.48550/arXiv.2401.04325.
- [246] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nusenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. doi:10.1109/CVPR42600.2020.01164.
- [247] N. Hurtós, X. Cufí, and J. Salvi. Calibration of optical camera coupled to acoustic multi-beam for underwater 3d scene reconstruction. In *OCEANS'10 IEEE SYDNEY*, pages 1–7, 2010. doi:10.1109/OCEANSSYD.2010.5603907.
- [248] Clayton Kunz and Hanumant Singh. Map building fusing acoustic and visual information using autonomous underwater vehicles. *Journal of Field Robotics*, 30(5):763–783, 2013. doi:10.1002/rob.21473.
- [249] Eduardo Iscar and Matthew Johnson-Roberson. Autonomous surface vehicle 3d seafloor reconstruction from monocular images and sonar data. In *OCEANS 2015 - MTS/IEEE Washington*, pages 1–6, 2015. doi:10.23919/OCEANS.2015.7404353.
- [250] Dushyant Rao, Mark De Deuge, Navid Nourani-Vatani, Stefan B Williams, and Oscar Pizarro. Multimodal learning and inference from visual and remotely sensed data. *The International Journal of Robotics Research*, 36(1):24–43, 2016. doi:10.1177/0278364916679892.
- [251] Monika Roznere and Alberto Quattrini Li. Underwater monocular image depth estimation using single-beam echosounder. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1785–1790, 2020. doi:10.1109/IROS45743.2020.9340919.

- [252] Haowen Liu, Monika Roznere, and Alberto Quattrini Li. Deep underwater monocular depth estimation with single-beam echosounder. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1090–1097, 2023. doi:10.1109/ICRA48891.2023.10161439.
- [253] Jiawei Zhang, Fenglei Han, Duanfeng Han, Jianfeng Yang, Wangyuan Zhao, and Hansheng Li. Integration of sonar and visual–inertial systems for slam in underwater environments. *IEEE Sensors Journal*, 24(10):16792–16804, 2024. doi:10.1109/JSEN.2024.3384301.
- [254] Luca Ebner, Gideon Billings, and Stefan Williams. Metrically scaled monocular depth estimation through sparse priors for underwater robots. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3751–3757, 2024. doi:10.1109/ICRA57147.2024.10611007.
- [255] Renato Silva, Aníbal Matos, and Andry Maykol Pinto. Multi-criteria metric to evaluate motion planners for underwater intervention. *Autonomous Robots*, 46(8):971–983, 2022. doi:10.1007/s10514-022-10060-x.
- [256] Muslikhin, Jenq-Ruey Horng, Szu-Yueh Yang, and Ming-Shyan Wang. Self-correction for eye-in-hand robotic grasping using action learning. *IEEE Access*, 9:156422–156436, 2021. doi:10.1109/ACCESS.2021.3129474.
- [257] Yiwen Chen, Sheng Guo, Zhiyang Liu, Lei Zhou, Zhengshen Zhang, Ruiteng Zhao, and Marcelo H. Ang. Economical precise manipulation and auto eye-hand coordination with binocular visual reinforcement learning. In *2024 9th International Conference on Control and Robotics Engineering (ICCRE)*, pages 6–11, 2024. doi:10.1109/ICCRE61448.2024.10589902.
- [258] Baichuan Huang, Jingjin Yu, and Siddarth Jain. Earl: Eye-on-hand reinforcement learner for dynamic grasping with active pose estimation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2963–2970, 2023. doi:10.1109/IROS55552.2023.10341988.
- [259] Xingjian Liu, Hari Krishnan Madhusudanan, Wenyuan Chen, Dahai Li, Ji Ge, Changhai Ru, and Yu Sun. Fast eye-in-hand 3-d scanner-robot calibration for low stitching errors. *IEEE Transactions on Industrial Electronics*, 68(9):8422–8432, 2021. doi:10.1109/TIE.2020.3009568.
- [260] Zhe Wang, Junfeng Fan, Fengshui Jing, Sai Deng, Mingyang Zheng, and Min Tan. An efficient calibration method of line structured light vision sensor in robotic eye-in-hand system. *IEEE Sensors Journal*, 20(11):6200–6208, 2020. doi:10.1109/JSEN.2020.2975538.
- [261] Mario Prats, J. Javier Fernandez, and Pedro J. Sanz. Combining template tracking and laser peak detection for 3D reconstruction and grasping in underwater environments. *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 106–112, 2012. doi:10.1109/IROS.2012.6385792.
- [262] Mario Prats, Javier Pérez, J. Javier Fernández, and Pedro J. Sanz. An open source tool for simulation and supervision of underwater intervention missions. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2577–2582, 2012. doi:10.1109/IROS.2012.6385788.

- [263] R. Conti, F. Fanelli, E. Meli, A. Ridolfi, and R. Costanzi. A free floating manipulation strategy for Autonomous Underwater Vehicles. *Robotics and Autonomous Systems*, 87:133–146, 2017. doi:10.1016/j.robot.2016.09.018.
- [264] Satja Sivčev, Matija Rossi, Joseph Coleman, Gerard Dooly, Edin Omerdić, and Daniel Toal. Fully automatic visual servoing control for work-class marine intervention rovs. *Control Engineering Practice*, 74:153–167, 2018. doi:10.1016/j.conengprac.2018.03.005.
- [265] Jiyong Li, Hai Huang, Yang Xu, Han Wu, and Lei Wan. Uncalibrated visual servoing for underwater vehicle manipulator systems with an eye in hand configuration camera. *Sensors*, 19(24):5469, 2019. doi:10.3390/s19245469.
- [266] Matheus Santos, Phillipe Santos, Petar Trsljic, Edin Omerdic, Daniel Toal, and Gerard Dooly. A monocular hybrid visual servoing controller for underwater inspection on work-class rovs. In *OCEANS 2021: San Diego – Porto*, pages 1–6, 2021. doi:10.23919/OCEANS44145.2021.9705705.
- [267] Jian Gao, Xiaomin Liang, Yimin Chen, Lijun Zhang, and Shisen Jia. Hierarchical image-based visual serving of underwater vehicle manipulator systems based on model predictive control and active disturbance rejection control. *Ocean Engineering*, 229:108814, 2021. doi:10.1016/j.oceaneng.2021.108814.
- [268] Roger Pi, Patryk Cieślak, Pere Ridao, and Pedro J. Sanz. Twinbot: Autonomous underwater cooperative transportation. *IEEE Access*, 9:37668–37684, 2021. doi:10.1109/ACCESS.2021.3063669.
- [269] Gideon Billings, Richard Camilli, and Matthew Johnson-Roberson. Hybrid visual slam for underwater vehicle manipulator systems. *IEEE Robotics and Automation Letters*, 7(3):6798–6805, 2022. doi:10.1109/LRA.2022.3176448.
- [270] Jiyong Li, Xinyu Bian, Hai Huang, Tao Jiang, Yihui Liu, and Lei Wan. Hybrid visual servoing control for underwater vehicle manipulator systems with multiple cameras. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(3):1742–1754, 2024. doi:10.1109/TSMC.2023.3329415.
- [271] Shan Qian and Y. Frank Cheng. Accelerated corrosion of pipeline steel and reduced cathodic protection effectiveness under direct current interference. *Construction and Building Materials*, 148:675–685, 2017. doi:10.1016/j.conbuildmat.2017.05.024.
- [272] Wanchai Yodsudjai and Thanawit Pattarakittam. Factors influencing half-cell potential measurement and its relationship with corrosion level. *Measurement*, 104:159–168, 2017. doi:10.1016/j.measurement.2017.03.027.
- [273] Ueli M. Angst. A Critical Review of the Science and Engineering of Cathodic Protection of Steel in Soil and Concrete. *Corrosion*, 75(12):1420–1433, 10 2019. doi:10.5006/3355.
- [274] Durgesh Tamhane, Sauvik Banerjee, and Siddharth Tallur. Monitoring corrosion in sacrificial anodes with pulsed eddy current and electromechanical impedance: A comparative analysis. *IEEE Sensors Journal*, 22(8):8147–8154, 2022. doi:10.1109/JSEN.2022.3157646.

- [275] W. H. A. Putra. Use of alternating current field measurement (acfm) technique for sizing internal surface defects on thin-walled austenitic stainless steel pipes. Master's thesis, Memorial University of Newfoundland, 1999. URL: <https://research.library.mun.ca/750/>.
- [276] Jaime Valls Miro, Nalika Ulapane, Lei Shi, Dave Hunt, and Michael Behrens. Robotic pipeline wall thickness evaluation for dense nondestructive testing inspection. *Journal of Field Robotics*, 35(8):1293–1310, 2018. doi:10.1002/rob.21828.
- [277] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer Cham, 2 edition, 2022. doi:10.1007/978-3-030-34372-9.
- [278] Mengkun She, David Nakath, Yifan Song, and Kevin Köser. Refractive geometry for underwater domes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:525–540, 1 2022. doi:10.1016/j.isprsjprs.2021.11.006.
- [279] Jorge J. Moré. The levenberg-marquardt algorithm: Implementation and theory. In G. A. Watson, editor, *Numerical Analysis*, pages 105–116, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg. doi:10.1007/BFb0067700.
- [280] Pedro Nuno Pereira, Raul Duarte Campilho, and Andry Maykol Pinto. Application of a design for excellence methodology for a wireless charger housing in underwater environments. *Machines*, 10(4):232, 2022. doi:10.3390/machines10040232.
- [281] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. doi:10.1145/358669.358692.
- [282] Nuno Abreu, Andry Pinto, Anibal Matos, and Miguel Pires. Construction progress monitoring – a virtual reality based platform. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6, 2023. doi:10.23919/CISTI58278.2023.10211855.
- [283] Siddharth Misra and Yaokun Wu. Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. In Siddharth Misra, Hao Li, and Jiabo He, editors, *Machine Learning for Subsurface Characterization*, pages 289–314. Gulf Professional Publishing, 2020. doi:10.1016/B978-0-12-817736-5.00010-7.
- [284] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016. doi:10.1109/cvpr.2016.90.
- [285] Laurent Sifre. *Rigid-Motion Scattering For Image Classification*. PhD thesis, Ecole Polytechnique, CMAP, 2014. URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.672.7091>.
- [286] Richard Hartley and Sing Bing Kang. Parameter-free radial distortion correction with center of distortion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1309–1321, 2007. doi:10.1109/TPAMI.2007.1147.

- [287] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. doi:10.1109/CVPR.2015.7298801.
- [288] Yotam Gil, Shay Elmaleh, Harel Haim, Emanuel Marom, and Raja Giryes. Monster: Awakening the mono in stereo, 2019. doi:10.48550/arXiv.1910.13708.
- [289] Andry Maykol Pinto. *Visual Motion Analysis based on a Robotic Moving System*. PhD thesis, Faculty of Engineering of the University of Porto, FEUP, 2014. URL: <https://hdl.handle.net/10216/73552>.
- [290] Carl Ludwig Siegel. A mean value theorem in geometry of numbers. *The Annals of Mathematics*, 46(2):340, 1945. doi:10.2307/1969027.
- [291] Miguel Fernando Paiva Velhote Correia. *Técnicas computacionais na percepção visual do movimento*. PhD thesis, Faculty of Engineering of the University of Porto, FEUP, 2001. URL: <http://hdl.handle.net/10216/11336>.
- [292] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 2015. doi:10.1016/j.cviu.2015.02.008.
- [293] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 402–419, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-58536-5_24.
- [294] Hamdan Al Jowair, Mansour Alsulaiman, and Ghulam Muhammad. Multi parallel u-net encoder network for effective polyp image segmentation. *Image and Vision Computing*, 137:104767, 2023. doi:doi.org/10.1016/j.imavis.2023.104767.
- [295] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 3–19, Cham, 2018. Springer International Publishing. doi:10.1007/978-3-030-01234-2_1.
- [296] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 12232–12241. IEEE, 2019. doi:10.1109/cvpr.2019.01252.
- [297] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, aug 2004. doi:10.1145/1015706.1015720.
- [298] Leslie N. Smith and Nicholay Topin. Super-convergence: very fast training of neural networks using large learning rates. In Tien Pham, editor, *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, SPIE, 2019. doi:10.1117/12.2520589.