FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Underwater Imaging and 3D Sensor Fusion

Diogo José Marques Silva

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Andry Maykol Pinto

July 29, 2024

Resumo

As limitações das atuais soluções de (operação e manutenção) O&M realizadas por humanos propõem a integração de tecnologias robóticas para enfrentar esses desafios. A presente dissertação analisa os desafios críticos da percepção subaquática no contexto da operação e manutenção (O&M) de energia eólica maritima. Ela destaca as dificuldades únicas da percepção subaquática, tais como a luz natural limitada, as condições imprevisíveis da água e o fenómeno de neve marinha. Ela explora a eficácia de vários sensores, incluindo sonares, câmeras e técnicas de imagem, como visão estereoscópica e Light Stripe Ranging (LSR), para melhorar a percepção subaquática. A visão estereoscópica é uma técnica de imagem que adquire informações 3D densas e coloridas, mas com limitada capacidade de estimação de profundidade. Por outro lado, o LSR é uma técnica de imagem capaz de fazer previsões precisas de profundidade, mas a densidade de informação adquirida é escassa. MARESye é um sensor híbrido que combina estas duas técnicas de imagem, permitindo a fusão destas técnicas de imagem. Ela discute a importância da fusão de dados heterogêneos em 3D na combinação de dados estereoscópicos e de LSR para permitir que robôs naveguem e operem de maneira eficaz em ambientes subaquáticos complexos. Em resumo, a dissertação sublinha a importância de sistemas avançados de percepção e fusão de dados 3D para melhorar a O&M da energia eólica maritima, reduzindo, em última análise, custos e aumentando a confiabilidade.

A rede AttentDeepUW é uma nova arquitetura de aprendizagem profunda projetada para melhorar a precisão e robustez da percepção subaquática por meio de mecanismos de atenção. Os resultados experimentais mostram um erro RMSE de 0,0167 m com dados sintéticos e uma precisão δ_1 de 99,1%. Em ambientes subaquáticos, a rede apresentou um erro absoluto médio de 0,0188 m e um erro relativo médio de 6.83 %. A previsão da rede é feita a partir de uma nuvem de pontos estéreo e uma nuvem de pontos LSR, que apresenta informações esparsas em apenas duas linhas. Esses fatores dificultam a estimativa, causando distorções nas previsões. No entanto, em cenários do mundo real, a rede gerou consistentemente nuvens de pontos de saída alinhadas e suaves. Estas experiências servem para validar a eficácia das metodologias propostas, apresentando melhorias substanciais na precisão da percepção e na eficiência operacional em comparação com os métodos convencionais. A rede opera com um tempo médio de processamento de 4,2 ms por iteração, enfatizando sua adequação para aplicações em tempo real onde tempos de processamento rápidos são essenciais.

O trabalho conclui destacando o impacto potencial desses sistemas avançados de percepção e técnicas de fusão de dados 3D na estrutura flutuante offshore DURIUS. Ao reduzir a dependência de mergulhadores humanos e ao melhorar as capacidades dos sistemas robóticos, a investigação visa reduzir os custos operacionais e aumentar a fiabilidade e segurança dos parques eólicos offshore.

Palavras-chave: O&M, visão 3D, subaquático, MARESye, fusão de dados heterogéneos, AttentDeepUW

Abstract

The limitations of current human-based (operation and maintenance) O&M solutions propose the integration of robotic technologies to address these challenges. The present dissertation analyses the critical challenges of underwater perception in the context of offshore wind energy operation and maintenance (O&M). It highlight the unique difficulties of underwater perception, such as limited natural light, unpredictable water conditions and the phenomenon of marine snow. It explores the effectiveness of various sensors, including sonars, cameras and imaging techniques such as Stereoscopc Vision and Light Stripe Ranging (LSR), to improve underwater perception. Stereoscopic Vision is an imaging technique that acquires dense and colorful 3D information but with limited capacity to depth estimation. On the other hand LSR is an imaging technique capable of making accurate depth predictions, but the information density acquired is sparse. MARESye is a hybrid sensor that combines these two imaging techniques, allowing the fusion of these imaging techniques. It discusses the importance of 3D heterogeneous data fusion in combining stereoscopic and LSR data to enable robots to navigate and operate effectively in complex underwater environments. In summary, the dissertation underlines the importance of advanced perception systems and 3D data fusion to improve the O&M of offshore wind energy, ultimately reducing costs and increasing reliability.

The AttentDeepUW network is a novel deep learning architecture designed to improve the accuracy and robustness of underwater perception through attention mechanisms. The experimental findings show an RMSE error of 0.0167 m with synthetic data and a δ_1 accuracy of 99.1%. In underwater environments, the network exhibited an average absolute error of 0.0188 m and an average relative error of 6.83 %. The network prediction is made from a input stereo point cloud and an LSR point cloud, which presents sparse information in just two lines. These factors hinder the estimation, causing distortions to the predictions. However, in real-world scenarios, the network consistently generated aligned and smooth output point clouds. These experiments serve to validate the efficacy of the proposed methodologies, showcasing substantial enhancements in perception accuracy and operational efficiency compared to conventional methods. The network operates with an average processing time of 4.2 ms per iteration, emphasizing its suitability for real-time applications where fast processing times are essential.

Work concludes by underscoring the potential impact of these advanced perception systems and 3D data fusion techniques on the DURIUS offshore floating structure. By reducing dependency on human divers and enhancing the capabilities of robotic systems, the research aims to lower operational costs and increase the reliability and safety of offshore wind farms.

Keywords: O&M, 3D vision, underwater, MARESye, heterogeneous data fusion, AttentDeepUW

SDG (Sustainable Development Goals)

This dissertation makes contributions to the fulfillment of the SDGs mentioned in table 1.

Table 1: Contributions of this dissertation to the SDGs.

SDG	**Target**	**Contribution**	**Performance indicators and metrics**
7	By 2030, increase substantially the share of renewable energy in the global energy mix.	By reducing maintenance time for offshore wind turbines, it is possible to increase the share of renewable energy in the global energy mix.	Renewable energy share in the total final energy consumption.
9	By 2030, upgrade infrastructure and retrofit industries to make them sustainable, with increased resource-use efficiency and greater adoption of clean and environmentally sound technologies and industrial processes, with all countries taking action in accordance with their respective capabilities.	Better offshore wind turbine maintenance systems promote greater use of this renewable energy source.	CO2 emission per unit of value added.
	Enhance scientific research, upgrade the technological capabilities of industrial sectors in all countries, in particular developing countries, including, by 2030, encouraging innovation and substantially increasing the number of research and development workers per 1 million people and public and private research and development spending.	This project is included in the highly scientific area promoting scientific development.	Research and development expenditure as a proportion of GDP.

Acknowledgements

Agradeço ao prof. Andry Pinto, por toda a ajuda e orientação prestada nesta longa dissertação. Agradeço sobretudo por todas as chamadas de atenção que permitiram guiar-me para um melhor desenvolvimento desta dissertação.

Agradeço imensamente ao Pedro Leite pelo constante acompanhamento e ajuda, pela paciência em auxiliar-me nas dificuldades encontradas ao longo do percurso, e pela sua prontidão em estar disponível para o que fosse necessário. A dedicação e apoio dele foram essenciais para a conclusão deste trabalho.

Um agradecimento especial ao Renato e ao Celso, que sempre se mostraram disponíveis para ajudar, mesmo não estando diretamente envolvidos com a minha dissertação. As contribuições e disposição deles em ajudar foram extremamente valiosas.

Por fim, gostaria de agradecer a todas as pessoas do laboratório por proporcionarem momentos de companheirismo e bons momentos durante esta fase. A convivência com todos tornou este percurso mais leve e enriquecedor.

Diogo José Marques Silva

"Education isn't something you can finish"

Isaac Asimov

Contents

List of Figures

List of Tables

Abbreviations

O&M	Operation and maintenance
SLAM	Simultaneous Localization and Mapping
3D	Tridimensional
2D	Two-dimensional
SIFT	Scale-Invariant Feature Transform
SURF	Speeded Up Robust Features
ORB	Oriented FAST and Rotated BRIEF
FAST	Features from Accelerated Segment Test
BRIEF	Binary Robust Independent Elementary Features
RANSAC	Random Sample Consensus
PS	Photogrammetric Stereo
LSR	Light stripe ranging
SNR	Signal-to-noise ratio
AUV	Autonomous underwater vehicle
TOF	Time-of-flight
LiDAR	Light Detection and Ranging
CNN	Convolutional Neural Network
USV	Unmanned Surface Vehicle
DOE	Diffractive Optical Element
ROV	Remotely Operated Vehicle
SGBM	Semi-Global Block Matching
RGB	Red Green Blue
CBAM	Convolutional Block Attention Module
SE	Squeeze-and-Excitation

Chapter 1

Introduction

1.1 Context and Motivation

Offshore wind production has seen a significant growth in recent years, becoming one of the main sources of electrical energy in the near future. Operation and maintenance (O&M) of offshore wind energy accounts for a large part of the costs of offshore wind energy produced. The underwater environment can cause deterioration in electrical and structural equipment, implying regular maintenance, and monitoring [1, 2]. Most of the current O&M solutions are man-based, which is a very arduous and dangerous process, requires a significant amount of resources (support vessels, human resources, etc.), and restricts operations to certain atmospheric conditions. Man-based O&M is also an unreliable and unrepeatable solution, with divers relying on hand-held flashlights to perform these operations. Employing robotic technologies in offshore wind farms can help to mitigate these issues [1]. Autonomous underwater vehicles, bridge substructures inspection systems [3], aerial mapping and localization [4], multi-domain inspection [5, 6], and docking procedures [7, 8, 9] are some examples where the implementation of robotics has become increasingly prevalent in these environments.
Developing advanced perception systems is now a pressing necessity in the field of robotics, as they are essential for enabling robots to navigate, interact, and operate efficiently in complex and dynamic environments [10]. Perceiving objects underwater is uniquely challenging when compared to other settings due to factors such as limited natural light, the absorption of light, unpredictable water conditions and disturbances in the environment that are beyond the control of human observers [10].

Sonars represent a category of acoustic sensors that exhibit resilience to underwater visibility challenges. These demonstrate robust performance even in the presence of water turbidity and are characterized by an extended range of detection capabilities, enabling them to navigate and perceive the underwater environment effectively. On the other hand, the acoustic waves used in sonar can result in lower detail and less precise imaging, making it challenging to recognize fine features or objects with intricate structures. Another limitation is its susceptibility to a minimum operational distance constraint. In the case of Infrared LiDAR, their utility is restricted when

deployed over extended distances beneath the water surface [10]. Specifically, water absorbs a significant portion, if not all, of the infrared laser energy emitted by these sensors. Consequently, the return signals become exceptionally weak or even non-existent, impeding the sensors' ability to provide reliable data over long underwater ranges. Imaging sensors, rely on optical sensors such as cameras can offer distinct advantages. The cameras employed in visual navigation systems take advantage of intricate and colorful information, proving instrumental in tasks related to object detection and classification within the underwater domain [10].

None of the existing sensors are capable of providing dense and very precise information about the environment. Stereoscopic Vision is an imaging technique that obtains 3D and colored information about the environment. Structured light is another imaging technique, which despite providing very sparse and non-colorized information about the environment, it presents high precision in depth estimation. Stereoscopic vision and structured light are two imaging techniques that have useful advantages, but each of them also has disadvantages that compromise their solo use. Stereoscopic vision provides dense, colorful 3D information, however with low depth accuracy. In contrast, structured light offers dense 3D information with high precision. MARESye is a sensor that explores these two sources of information. As these disadvantages are complementary, it becomes theoretically possible to implement a heterogeneous data fusion to minimize the individual disadvantages of each one [11].

1.2 Objectives

The objective of this dissertation is to provide a high-density and precise 3D reconstruction of the underwater environment. Data fusion is the process of combining data from multiple image techniques to create a more comprehensive understanding of the environment. By integrating data from various techniques, a robot can gain a more accurate and holistic perception of its surroundings. Passive Photometric Stereo (PS) and active Light Stripe Ranging (LSR) techniques will be combined to produce a high-density and precise 3D reconstruction of the environment. In summary, the dissertation aims to:

- Develop 3D heterogeneous data fusion algorithms based on deep learning for integrating visual information (that are textured and dense information) and precise depth information (provided by LSR technique that is sparse but very accurate information) from the MARESye hybrid sensor into a unified 3D point cloud providing a better reconstruction of the underwater environment.

- Evaluate data fusion algorithms by conducting initial validation of performance using synthetic data. This involves testing the algorithms with synthetic data and then applying them to real-world data to assess their effectiveness and reliability.

- Assess the performance of data fusion algorithms using data collected in a controlled underwater environment, ensuring accurate and consistent results under specified conditions, improving the state-of-the-art using attention mechanisms.

1.3 Document Structure

In addition to this introductory chapter, the document presents other four other chapters. Chapter 2, presents the state of the art, including section 2.1, that presents the challenges of the underwater environment which must be overcome by robotic implementations. The section 2.2 presents sensors that provide a perception of the surrounding environment and also some image techniques that provide processing and enhancement of information provided by the sensors. The final section of the chapter (2.4) provides a brief review of the most relevant information of the state-of-art. Chapter 3 begins with a detailed presentation of the sensor used and a comprehensive characterization of the problem. It includes a description of data acquisition procedures and a detailed overview of the final deep learning network developed. The analysis of the obtained results is presented in Chapter 4. Chapter 5 provides a summary of the obtained conclusions, as well as future work to be developed.

Chapter 2

Bibliographic review

This chapter focuses on the main challenges of the underwater environment and presents sensors, imaging techniques, heterogeneous 3D data fusion techniques.

2.1 Challenges of underwater perception

One of the most significant challenges in underwater perception is the scarcity of high-quality sensors capable of delivering tridimensional data. Obtaining such information is impeded by a range of factors, including limitations in sensor range and resolution, suboptimal lighting conditions, and the absence of textural cues in the environment [12].

Another challenge is optical sensor calibration issues and lens distortions. When light rays enter a spherical lens, they are refracted or reflected more or less than those that strike close to the center. This deviation reduces the quality of images produced by optical systems. The spherical shape of a lens can cause light rays to deviate and not focus at a single point, leading to image distortion. This is particularly noticeable when the lens is large or has a short focal length [10].

2.1.1 Absorption

The underwater environment exhibits significant light absorption, primarily due to interactions between photons and water molecules, resulting in heat generation. This interaction hinders the penetration of visible light into water, particularly affecting red and violet spectrums, causing a substantial decrease in their intensity. The impact is less pronounced in the blue and green range. As a result of this absorption effect, red and violet light vanishes at depths less than 7 m, yellow and orange colors are lost around 15 m, and blue and green colors fade at approximately 30 m, as seen in figure 2.1 [13]. This phenomenon explains the prevalent blue or green appearance of seawater. Furthermore, underwater substances like organic matter and dissolved organic materials can intensify light attenuation, especially in the blue wavelength. Other factors contributing to reduced ambient light in underwater settings include light reflection at the water's surface, light refraction as light traverses the water, and light diffusion causing deviations in light paths. These

interactions collectively restrict the penetration of visible light into the water, resulting in darker surroundings as a robot descends into deeper waters [13].

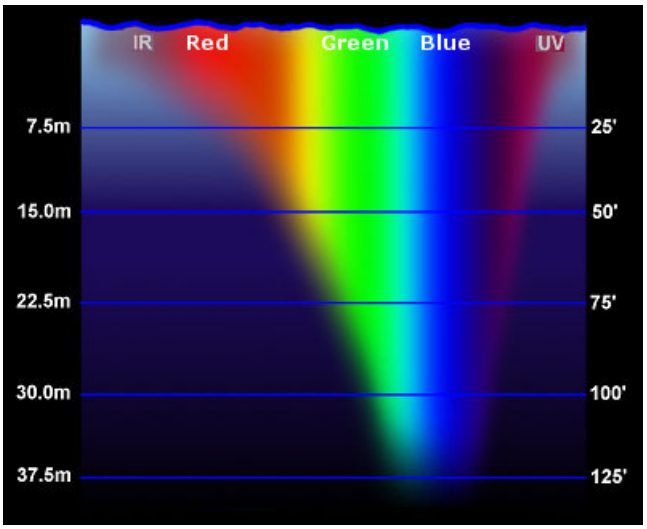


Figure 2.1: Attenuation of light in water[1].

To address the issue of low visibility in underwater environments, active light systems integrated with robotic platforms are used. However, this solution faces two main challenges. First, the artificial light system also experiences energy attenuation, limiting its effectiveness in illuminating deeper water layers.

While increasing the overall object illumination might seem beneficial, the back-scattering effect diminishes the contrast between the object and its background. The problem stems from light being scattered along the illumination path, washing out the object and causing light from the object to scatter, resulting in image blurring. Enhancing the total object illumination doesn't improve contrast in such scenarios because the scattering is directly proportional to intensity, negating any net increase in contrast. In turbid environments, these perception challenges worsen as higher turbidity amplifies light attenuation and scattering intensity, particularly in waters rich in clay, silt, algae, and other organic matter, rendering optical imaging systems ineffective[13].

2.1.2 Scattering

One other challenge of underwater perception is the scattering phenomenon. The interaction of light with water molecules and suspended particles significantly influences the propagation of light rays. Scattering refers to the phenomenon where light gets dispersed in various directions due to interactions with particles in the water. This scattering of light contributes to the degradation of image clarity, resulting in low contrast and blurry details. The scattering and absorption of light in water can cause various quality degradation issues in underwater images. These degraded-quality underwater images are harmful to analysis and applications [14, 15].

[1] https://www.empiricalimaging.com/knowledge-base/underwater-photography/

Backscattering occurs when light is redirected backward after encountering particles or interfaces within a medium. This phenomenon plays a significant role in various fields, including remote sensing, medical imaging, and underwater exploration, influencing the interpretation of collected data and the quality of acquired images. Back-scattering frequently arises in underwater imaging when the active light system inadvertently illuminates particles between the optical sensor and the object or the open water space behind the object rather than the intended object itself. This leads to blurred and noisy images in various underwater optical systems. This effect restricts object detection distance in contrast-limited imaging applications like human vision or film [14].

Marine snow represents a significant source of degradation in underwater images (see figure 2.2), and it's a common occurrence in the ocean. In the context of capturing underwater images or videos, these small particles intercept the path of light before it reaches the optical sensor. As a result, these particles represent an effect of backscattering the light, leading to issues such as reduced contrast and a hazy appearance. Additionally, light reflecting off marine snow particles introduces random bright spots in the captured images, diminishing the overall clarity of the scene [16].

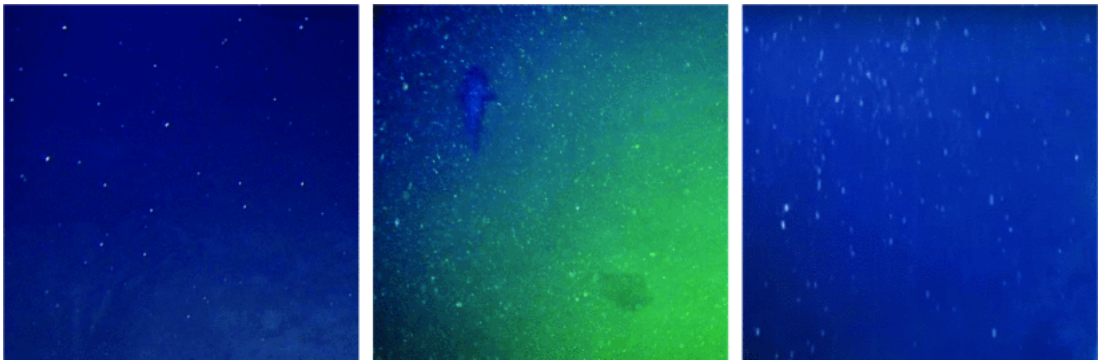


Figure 2.2: Marine Snow in underwater environment [16].

The irradiance $E(r)$ at position r can be described by equation 2.1, where a and b are absorption and scattering coefficients [11].

$$E(r) = E(0)e^{-ar}e^{-br} \tag{2.1}$$

The underwater medium, characterized by its distinct optical properties, introduces challenges and intricacies that profoundly impact the way objects and features are perceived. Understanding the principles of scattering is essential for developing effective underwater sensing and imaging systems.

To effectively solve these quality degradation issues, various methods have been introduced. One of these methods is an underwater image restoration method via weighted wavelet visual perception fusion. This method first presents an attenuation-map-guided color correction strategy to correct the color distortion of an underwater image. Then it employs strategies to obtain global and local contrast-enhanced images. Finally, it introduces a weighted wavelet visual perception

fusion strategy to obtain a high-quality underwater image by fusing the high-frequency and low-frequency components of images at different scales. These methods aim to improve the quality of underwater images, making them more suitable for human perception and computer processing [14].

2.2 Underwater Perception: Sensors and Imaging Techniques

2.2.1 Sonars

Sonar systems work by emitting a sound wave and then listening for the echo. This is known as the "time of flight" method. The time it takes for the echo to return can be used to calculate the distance to an object. Sonars are a type of sensor that has the advantage of not suffering from the underwater conditions imposed on optical sensors. They are robust to water turbidity, have long range due to the sound properties of water, and are immune to lighting conditions. However, sonars have specific disadvantages [10]. However, at very short distances, the echo may return while the sonar system is still emitting the sound wave. This can cause a problem because the sonar system needs to switch from emitting to receiving mode to detect the echo. If the echo returns before the system has switched to receiving mode, it won't be detected, and the sonar system will not be able to accurately measure the distance to the object. This system is proposed for longer ranges and is unsuitable for short-range applications. Additionally, the utilization of sonars is hindered by their suboptimal resolution, a critical limitation in inspection and maintenance tasks where precise millimeter-level reconstruction is essential [2].

Historically used in underwater applications, sonar-based systems have made it possible to carry out tasks such as bathymetry, navigation, and collision avoidance, as the work proposed by Y. Petillot *et al.* (2001) [17], where a new structure was designed for sonar image segmentation, underwater object tracking, and movement estimation. The maintenance and control of port structures used to be carried out by divers. This type of human operation entails various costs and risks, which is why N. Brahim *et al.* (2008) [18] proposed a sonar-based system for inspecting quays. The work proposes to detect and characterize quay defects using sonar images.

P. Teixeira *et al.* (2016) [19] proposed a submap-based technique for inspecting and mapping underwater structures with complex geometries. The approach is based on the use of probabilistic volumetric techniques creating submaps from multibeam sonar scans. A slightly different approach Y. Kim *et al.* (2020) [20] use a sonar on an Unmanned Surface Vehicle (USV). Because the sonar is installed on a surface vehicle, the waves affect the sonar data. The author also proposes a stabilization method to minimize image errors. T. Guerneve *et al.* (2015) [21] propose an underwater 3D reconstruction solution based on 2D imaging sonars. This algorithm generates 3D maps based on a sequence of imaging sonar images. This technique allows surface reconstruction for tasks of inspection using standard sonars. S. Hou *et al.* (2022) [3] used a sonar and a convolutional neural network to inspect the underwater part of bridges. The convolutional neural network

(CNN) provides quantitative measurements of erosion depths and damage using the data obtained from the side-scan sonar device.

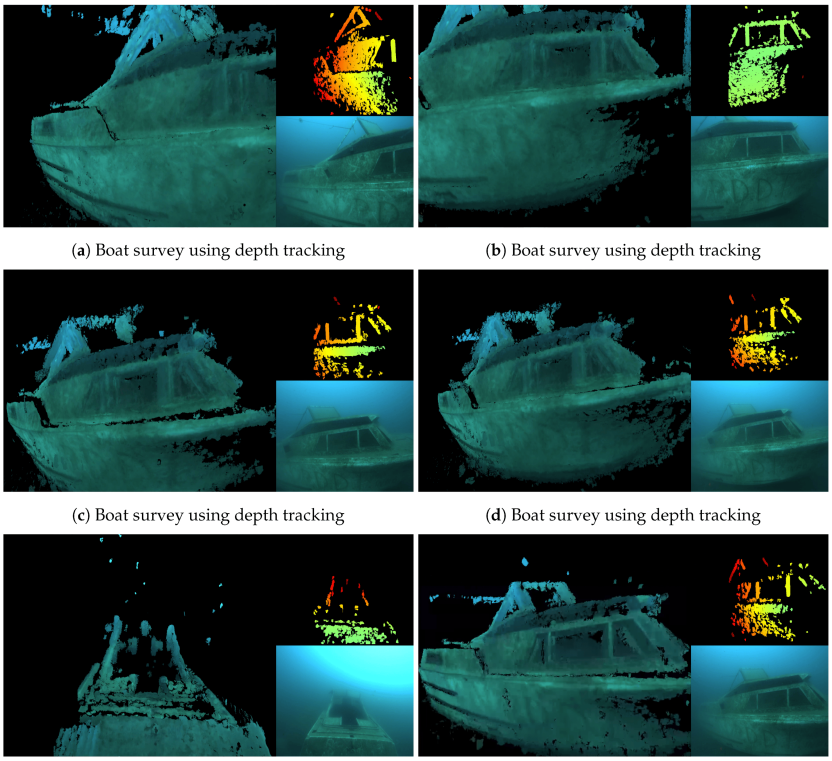
2.2.2 Imaging Techniques

A frequently employed method for underwater 3D reconstruction involves the utilization of optical systems that have the advantage of providing dense, extensive information about the environment, and unlike sonars, they provide rich color and texture information supporting a large range of tasks [10]. They are passive sensors and therefore, in low light conditions, which are very common in underwater environments, they require lighting on board the vehicle to be useful. Furthermore, in conditions of low visibility and high water turbidity, optical sensors may not be useful underwater perception. These water conditions can corrupt optical sensor data. Another drawback is the fact that the quality of water causes a heavy attenuation of the red channel and haze the images, reducing the texture of images which is fundamental for automated perception methods [10]. Imaging techniques fall into 2 categories: active and passive [22]. Passive underwater imaging uses external light sources (natural or artificial) to capture different points of view of the environment to obtain information for 3D reconstruction. This type of perception system is usually based on stereo optical sensor pairs [2, 22]. Contrarily, active underwater imaging consists of projecting signals, such as waves, pulses, lasers, or light patterns into the environment and then detecting and analysing them [22]. Tridimensional data extraction is achieved by employing triangulation methods.

Stereo Vision

Stereo vision is a classic computer vision algorithm inspired by the human binocular vision system. It relies on two parallel viewpoints and calculates depth by estimating disparities between matching key-points in the left and right images. Incorporating multiple visual sensors into a system enables the perception of depth and the generation of a tridimentional (3D) representation of the surrounding environment. This method involves the use of multiple optical sensors to capture several images of an object, with the optical sensors positioned at a known distance from each other (displacement), a discernible disparity among objects within captured images can be established. This inter-sensor disparity serves as the basis for estimating the depth of objects in the visual images. The integration of these images relies on triangulating the distances between the optical sensors and the distance from the scanned object, ultimately generating a 3D image. However, achieving precise image matching poses challenges, needing the implementation of specific procedures to ensure accurate results [23]. The figure 2.3 shows a 3D reconstruction from stereo matching.

The implementation of 3D image sensing in various industries serves underwater purposes like underwater navigation and collision detection systems. Different methods are employed for 3D imaging depending on the specific application requirements. While methods like laser point are



(a) Boat survey using depth tracking (b) Boat survey using depth tracking

(c) Boat survey using depth tracking (d) Boat survey using depth tracking

(e) Boat survey using depth tracking (f) Boat survey using colour tracking

Figure 2.3: Stereo Reconstruction of a boat in underwater [24].

utilized for accurate reconstruction, they may not be suitable for capturing dense, dynamic scenes, a capability achieved by stereo vision [23].

Utilizing the principles of triangulation, depth can be derived through the application of similar triangles. Considering the left image as the reference, equations 2.2 and 2.3 are formulated by accounting for the disparity induced in the right image due to its displacement [25].

In this context, where X and Z denote the lateral distance and depth of the object relative to the optical sensor, and x^L and x^R represent the x-coordinates of pixels in the left and right images, respectively, with f representing the focal length, the following relationships are established.

$$\frac{Z}{f} = \frac{X}{x^L} \tag{2.2}$$

$$\frac{Z}{f} = \frac{X-T}{x^R} \tag{2.3}$$

$$d = f\frac{T}{Z} \tag{2.4}$$

The stereo vision system typically comprises two optical sensors with identical specifications, including the same focal length, aperture, and sensor area. Ideally, the left and right optical sensors are aligned in the same plane, ensuring that their horizontal axes are on the same line and parallel

to the imaging plane. The imaging model is depicted in Figure 2.4, with Figure 2.5 illustrating the two-dimensional plane of the ideal model [25].

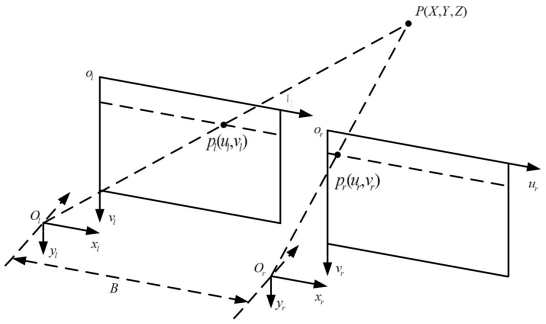


Figure 2.4: Representation of a point in 3D space in stereo system [25].

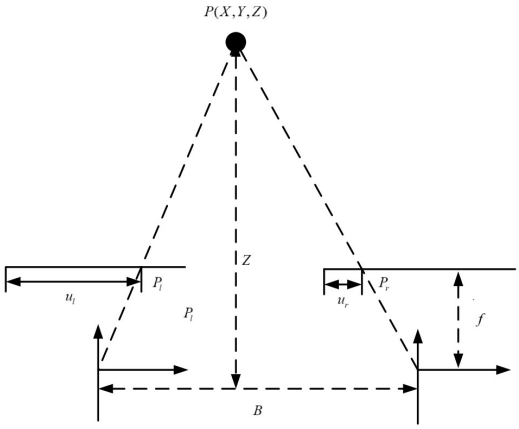


Figure 2.5: 2D perspective of stereo system operation [25].

In the figure 2.4, the distance between the two optical sensors is denoted as B. The point P(X, Y, Z) in the imaging coordinates of the left and right optical sensors is represented as $p_L(u_L, v_L)$ and $p_R(r_R, v_R)$ respectively. The left and right optical sensor coordinate systems are respectively displayed as $O_L x_L y_L$ and $O_R x_R y_R$. The image coordinates of the left and right optical sensors are shown as (u_L, v_L) and (u_R, v_R). The conversion from a tridimensional map to a two-dimensional map is illustrated in Figure 2.5.

In the realm of stereo vision, a fundamental relationship exists between depth and disparity. This association is characterized by an inverse proportionality: an increase in disparity corresponds to a closer positioning of an object to the optical sensor baseline, whereas a decrease in disparity indicates a greater distance from the baseline. As explained in figure 2.6, the disparity observed in stereo images is directly linked to the baseline between the two optical sensors. When the baseline is diminished, the resulting disparity is likewise reduced, yielding smaller differences between the images. Conversely, an augmentation of the baseline leads to a proportional escalation in disparity. These principles bear significant implications for the design of stereo vision systems. In the pursuit of accurate depth measurement, precise disparity assessment becomes paramount. Consequently, an optimal stereo configuration necessitates a sufficiently large baseline, as an expanded baseline facilitates more meticulous disparity measurements. This understanding underscores the critical importance of thoughtful design considerations when developing stereo systems for applications where precise depth estimation is imperative [26].

In recent years many research works rely on underwater perception systems based on optical stereo systems. M. R. Shortis *et al.* (2014) [28] introduced stereo-video system to towed body systems. In the same year, K. Williams *et al.* (2014) [29] deployed underwater stereo optical sensor capable of triggering when animals are present in the field of view. P. Carrasco *et al.* (2015) [30] propose a stereo-vision Graph-SLAM system using a conventional Bumblebee stereo pair to

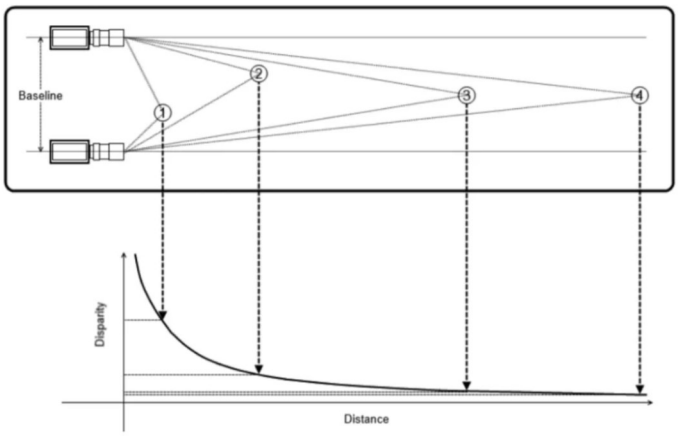


Figure 2.6: Accuracy of depth estimation by distance [27].

be used in control and navigation systems of SPARUS II AUV, and M. Carreras *et al.* (2018) [31] continues the project proposing a path-planning algorithm. A stereovision system composed of two AVT Mako optical sensors to 3D object detection was proposed F. Oleari *et al.* (2015) [32] to provide visual perception to underwater intervention and manipulation tasks. S. Tani *et al.* (2023) [33], develop a navigation solution based on stereo vision. Vision systems were used to collect images of the underwater environment performing robot navigation based on visual odometry. V. Kramar *et al.* (2023) [34] investigate methods to tackle challenges associated with detecting, recognizing, and localizing objects in underwater environments, employing stereo vision systems to overcome environmental constraints. The analysis focuses on the technical limitations presented by underwater conditions and the effectiveness of stereo vision systems in addressing these issues.

Structured Light

Light Stripe Ranging (LSR) is a technique used in imaging systems, particularly in underwater environments that involves projecting a set of visible stripes of light into the scene and recovering 3D information from these laser stripes through triangulation. It is one of the sensors/techniques used to gather sparse depth information for dense disparity maps from RGB and sparse depth information using deep regression models [2]. It has several advantages and disadvantages: Advantages include giving accurate 3D information (LSR provides accurate 3D information from harsh underwater environments) and less affected by sub-sea conditions. Disadvantages include limited data acquisition, only providing 3D information in a narrow line.

To extract correct 3D information from lasers, it is necessary to conduct a calibration to determine the spatial configuration of each laser in relation to the camera frame. Triangulation calculates the tridimensional points by intersecting 2D points derived from segmentation with the given plane equations (assuming the camera matrix is known). Equations 2.5, 2.6, 2.7 represent the triangulation to calibrate the lasers.

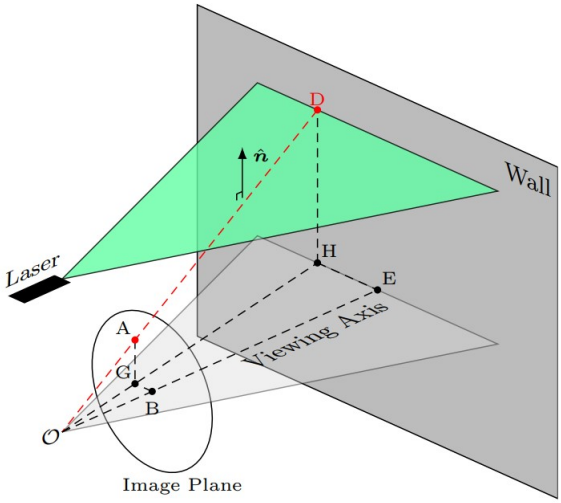


Figure 2.7: Depth estimation by triangulation between optical sensor and laser projection [35].

$$Z = -\frac{a*x+b*y+d}{c} \tag{2.5}$$

$$X = Z*x \tag{2.6}$$

$$Y = Z*y \tag{2.7}$$

G. Inglis *et al.* (2012) [36] proposed a structured light laser imaging (as demonstrated in figure 2.8) to create high-resolution bathymetric maps of the sea floor.

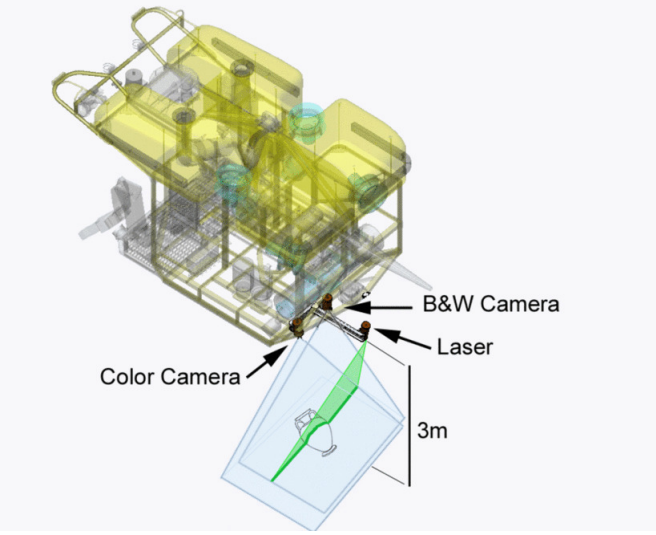


Figure 2.8: Bathymetry using the LSR system proposed by G. Inglis *et al.* (2012) [36].

N. Hansen *et al.* (2015) [35] proposed the use of two lasers and a optical sensor, the lasers

project vertical lines into the environment parallel to the optical sensor's axis of view. By triangulating points in the image (figure 2.7) using the Hough transform, this system provides an accurate estimate of depth and is used in underwater inspection by autonomous underwater vehicles (AUVs). This system showed an approximated error of 0.04m in the depth measurements in underwater inspection tasks. Although these solutions show good accuracy in depth measurements, they reveal a sparse point cloud. A solution was developed by F. Lopes *et al.*(2015) [37] use a rotating laser line projector mechanism that allows an area to be scanned with a single beam. This system is more complex and difficult to set up, but it allows for denser reconstruction and is capable of reconstructing an object with an overall error of 2% of its size. The paper also shows how to calibrate this stereo SLS system in and out of water and is able to reveal accuracy results of less than 1 mm in a dry environment. M. Massot-Campos *et al.* (2014) [38] proposed a different approach. Instead of increasing the number of laser beams, the proposed system uses a Diffractive Optical Element (DOE) in front of the laser beam, dividing it into 25 parallel lines. This strategy allows for a more complete 3D reconstruction without increasing the number of lasers and restricting energy consumption. However, this approach is highly affected by the turbidity of the water as the refracted beams have less energy and are quickly attenuated at short distances.

Usually, underwater 3D laser scanners rely on a rotating mirror driven by a galvanometer. However, the planes of light directed by these mirrors are usually deformed into cones. For this reason, M. Castillon *et al.* (2021) [39] proposed the use of a biaxial MEMS mirror, in which the second rotational degree of freedom can be used so that the refraction process transforms the light shapes into planes. Y. Ou *et al.* (2023) [40] addressed active vision measurement systems designed for underwater 3D reconstruction based on binocular structured light to combat the challenges of light scarcity in underwater operations. H. Lin *et al.* (2024) [41] developed a high-precision 3D reconstruction method for underwater concrete using line-structured light combined with stereo vision. This method features a mathematical model to address light refraction, utilizes epipolar constraints for noise reduction, and employs dual cameras for enhanced color accuracy, achieving less than 5% error in controlled tests.

Pattern Projectors

Pattern projector is a technique based on designing a pattern or sequence of patterns that uniquely determines the keyword of a pixel within a non-periodic region (each point on the surface of the object has a unique binary code that differs from the code of any other point) [42, 43]. The 3D coordinates of each point can be calculated based on triangulation principles.

F. Bruno *et al.* (2021) [44] use a gray-coded pattern projector and stereo equipment to reconstruct submerged 3D objects. This system was difficult to operate at high levels of water turbidity, making it unable to detect the projected patterns. It also had high acquisition times which made it impossible to optimize the system's performance. A different approach was presented by A. Sarafraz *et al.* (2016) [45]. In this approach, only a single optical sensor is placed underwater and an out-of-water pattern projector is used, projecting from top to bottom. This approach has several disadvantages being highly dependent on the clarity of the water and given the strong attenuation

of the projected light its operating range is very limited. Another similar work was developed by Q. Zhang *et al.* (2011) [46] where a fringe projector and an optical sensor were positioned out of the water and the object to be tracked was in the water. The digital projector projects a sinusoidal fringe and the optical sensor records the distorted fringe which is modulated by the shape of the object. The effects of the air-water interface were also taken into account. S. Zhuang *et al.* (2023) [47] used a pattern projector to overcome the challenges posed by the underwater environment. It used an active speckle pattern method in conjunction with underwater stereo vision to improve the accuracy of underwater 3D measurement. This type of binary coding is reliable and less sensitive to surface characteristics since all pixels are coded by binary values. However, to achieve high spatial resolution, it is necessary to design a large number of sequential patterns. For the application to be developed, 3D image acquisition would be very high, making it unfeasible to use [43]. Underwater 3D reconstruction is challenged by equipment nonlinearities and varied HDR object reflectance, causing phase errors. Z. Zhu *et al.* (2023) [48] propose a double N-step orthogonal polarization state phase-shift strategy (DOPS), using orthogonal polarization to enhance phase accuracy and efficiency. Experiments show DOPS reduces errors by 57% and increases efficiency by 50% compared to existing methods.

2.2.3 Hybrid Sensors

A lot of research has already been done on underwater perception systems. However, the challenges posed by underwater environments often do not allow for the millimetric precision required to perform minute tasks due to data degradation. A hybrid sensor that combines active and passive sensing the system can exploit the benefits of each mode and overcome the limitations of the other. For example, active mode can provide high resolution and accuracy, but it also consumes more power and generates more interference. Passive mode can provide low power consumption and textured information.

Given the advantages of stereo and LSR systems, it is useful in this work to use a sensor that allows the use of these imaging techniques and is capable of being used in an underwater environment. MARESye (figure 2.9) is a hybrid image sensor that can be easily installed in different underwater robotic applications. MARESye provides dense and accurate 3D information of diverse underwater environments. The system is guided by a range-gated system to reduce the impact of photometric problems such as diffuse reflection, non-uniform illumination, and water turbidity. This system can be easily installed in different robotic applications, being a self-sufficient system with an internal processing unit. MARESye ensures high fidelity of the data retrieved and has a built-in information fusion module that allows for a dense and accurate tridimensional representation of the data obtained [11].

2.3 Heterogeneous 3D Data Fusion

Measurements obtained by sensors are always prone to errors, with noise introduced into the measurements due to phenomena imposed by the environment. Fusing information from different

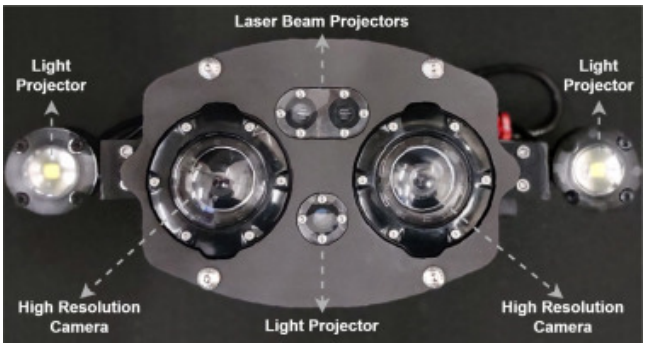


Figure 2.9: MARESye hybrid imaging sensor [2].

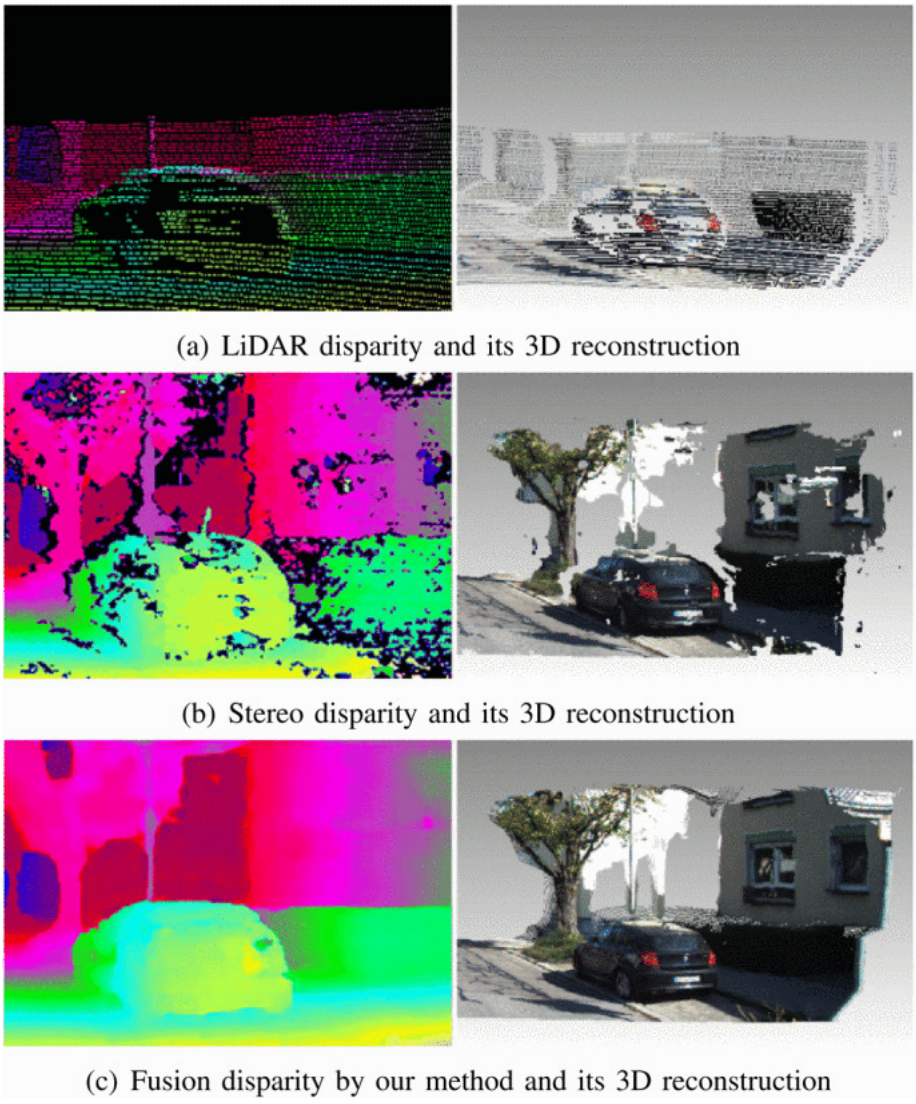
sources becomes a multidisciplinary solution to combat the weaknesses of each sensor or technique [49].

In an out-of-water context, a lot of research has already been done into fusing heterogeneous tridimensional information from stereo vision and LiDAR data. Stereo data is textured and dense but requires a lot of computing power to produce accurate results [50]. On the other hand, LiDAR data can be sparse or dense and is very accurate. This complementarity makes the fusion of these data useful and feasible, and W. Maddern has developed work in this area. W. Maddern *et al.* (2016) [50] fused sparse 3D data from a LiDAR scanner, which produces accurate but low-density depth maps, with stereo data. The authors propose a real-time probabilistic approach that merges the depth maps by propagating uncertainty estimates through a prior disparity refinement phase. This system can be used in localization, mapping, and collision avoidance tasks for autonomous vehicles. The method was evaluated on data collected by small urban autonomous vehicles and made use of the KITTI dataset.

K. Park *et al.* (2018) [51], presented a new approach to fusing the same type of data mentioned above. K. Park *et al.* chose to use a deep convolutional neural network (CNN) architecture for high-precision depth estimation. In this network, the complementary characteristics of sparse 3D LiDAR data and dense stereo depth are coded simultaneously in an enhancing way, differing from other CNNs by incorporating a compact convolution module. The authors report accurate results, as evidenced by an example in figure 2.10, on several data sets proving the generalization capabilities of the proposed network.

D. Martins *et al.* (2018) [52] used a self-supervised approach in which the depth estimates obtained from the stereo data are used in a convolutional neural network (CNN), transforming a single fixed image into a dense depth map. After training, the monocular estimates obtained from the CNN and the stereo estimates are fused to preserve the high reliability of the stereo and take advantage of the monocular depth in occluded regions. The experiments use the KITTI dataset and aim to show that this type of fusion leads to better performance than isolated stereo estimates.

Using stereo optical sensors providing dense data and laser ranging sensors being more precise and sparse, M. K. Ali *et al.* (2019) [53] proposed a new mechanism for incrementally merging sparse data with dense data to produce dense and accurate depth maps. This method proved to produce better results than those from a single source.



(a) LiDAR disparity and its 3D reconstruction

(b) Stereo disparity and its 3D reconstruction

(c) Fusion disparity by our method and its 3D reconstruction

Figure 2.10: Example of improving the quality of reconstruction after fusion [51].

An unsupervised LiDAR-stereo fusion approach that can be trained end-to-end without the need for real data by using a feedback loop to confirm the data. This approach by the authors X. Cheng *et al.* (2019) [54] overcome the lack of LiDAR information.

Another approach is to fuse two equally dense sources based on a one-to-one mapping of 3D points. This approach is usually used in aerial robotic applications, but in underwater environments, it is difficult to use this approach due to the lack of dense and precise information, such as that provided by LiDARs, which makes it impossible to use in underwater environments due to the strong attenuation of red and infrared radiation.

Recently there was a scientific advance with the work developed by P. N. Leite *et al.* (2024) [2] bring for the first time a 3D information fusion approach capable of being used in the inspection and maintenance operations of submerged structures. This approach relies on dense information with textures and more accurate and sparse data, triangulate the projection of laser beams in the scene. This approach contrasts with other fusion approaches, in which case the sparse 3D information is propagated to the dense point cloud by exploiting homogeneous regions around the specific beams. This way, an equally dense input cloud is no longer necessary to serve as a reference. A supervised learning approach, called RHEA, is also discussed, which is based on state-of-the-art approaches for training models with synthetic data, however, a synthetic-to-real training scheme is used to allow direct application in an underwater context, skipping a retraining phase.

J. Zhang *et al.* (2014) [55] create a perception system using sonar and stereo vision to be integrated into an ROV. A fusion of these two perception methods is proposed to obtain an improved perception system for maintaining underwater infrastructures.

2.4 Critical Review

This section presents a short analysis of the drawn conclusions concerning the different state-of-art sensors and image techniques used and described in this chapter. Although sonars are robust and unaffected by water turbidity, which is why they are widely used in underwater environments, they lack the resolution required for inspection tasks. Another factor that makes it unsuitable for these tasks is its high minimum usage distance, making it unsuitable for short-distance tasks. Visual sensors, on the other hand, have excellent resolution and textured information, making them an acceptable choice for the task at hand. However, to obtain 3D information using visual sensors, imaging techniques will have to be applied, stereo vision being one way of obtaining dense and textured information about the scene, although the accuracy of the estimated depth is not the best. To obtain more precise depth information, techniques such as LSR could be used, projecting lasers onto the scene. This technique provides accurate depth information, but the information extracted is sparse and untextured. Structured light (Pattern Projector) is another of the imaging techniques covered in this document. The advantage of using a hybrid sensor that combines stereo vision and LSR is that it can improve the accuracy and robustness of the 3D reconstruction. Stereo vision can provide high-resolution images and dense depth maps, but it also suffers from occlusion, noise, and low-texture regions. LSR can provide accurate and reliable distance measurements, but it

also has a limited field of view and resolution. By fusing the data from both sensors, the hybrid sensor can overcome the limitations of each sensor and provide a more complete and consistent 3D model of the scene. To collect this data, the MARESye hybrid sensor is prepared to use these two techniques and is ready to be used in an underwater environment.

Chapter 3

Fusing heterogeneous 3D information

Light propagation is challenged by adverse underwater physical phenomena that affect light propagation through water. Absorption and scattering, as the main challenging factors, degrade the information obtained from the environment.

The objective of this chapter is to presents a AI model that is to takes advantage of these two imaging techniques provides a dense and precise unified point cloud on the information obtained by the MARESye sensor. The system must be capable of delivering an information density comparable to that of stereo, ensuring that each point is corrected using the sparse information from the LSR. This approach ensures that all points in the dense point cloud are refined using the limited yet precise information obtained from the sparse point cloud. Therefore, the depth information from the LSR will be propagated by the stereoscopic points to correct its depth values. The integration of data will be executed by projecting the point clouds onto the camera's reference plane, thereby converting the point clouds into 2D images. These images will subsequently be processed through a deep learning network to generate a unified 2D representation. The deep learning network employs a U-Net architecture [56], operating in an early-fusion process, utilizing a ResNet18 backbone that has undergone pre-training for the encoder component. Additionally, attention blocks are implemented within the skip connections to enhance the emphasis on the most significant features.

This chapter explores the data acquisition process (section 3.1) and the augmetations used (subsection 3.1.1), the architecture of AttentDeepUW and AttentDeepUW_3skips (section 3.2) and the network optimization methods (subsection 3.2.1).

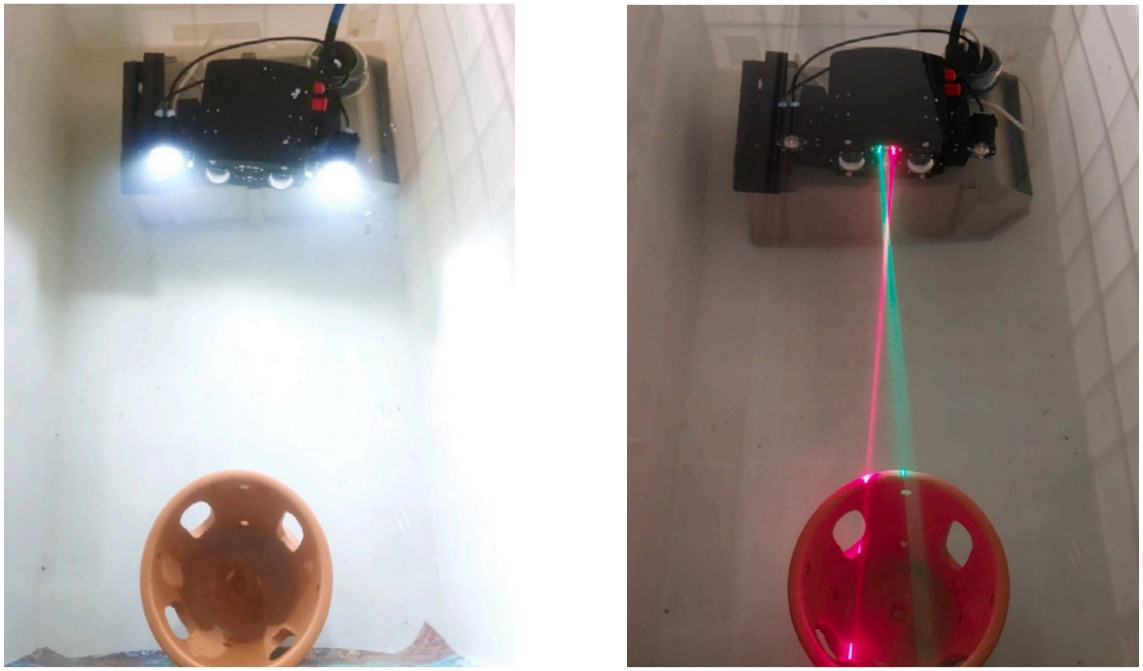
3.1 Underwater Tridimensional Data Acquisition

MARESye is a hybrid sensor [1] equipped with a dual-camera system, each camera featuring a resolution of 1440 x 1080 pixels arranged in a stereo configuration. The device also includes two

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

laser beam projectors operating at red and green wavelengths, complemented by a set of high-intensity LED lights. An internal processing unit facilitates autonomous data processing, avoiding the need for external computational support. Additionally, the sensor is equipped with a trigger system. This arrangement improves image clarity and precision. Additionally, the sensor's compact form factor facilitates integration into Autonomous Underwater Vehicles (AUVs), Remotely Operated Vehicles (ROVs), and robotic manipulators. It is engineered to endure pressures encountered at depths of up to 300 m. The baseline of the optical sensors and the subtle orientation of the LED lights are precisely calibrated to achieve an optimal working range of 0.5 to 0.8 m. Operation beyond this designated range is possible, it may adversely affect the quality of the data collected [11, 2].

As MARESye is a hybrid sensor, the sensor is capable of obtaining information from the environment through active and passive techniques as shown in figure 3.1. These different 3D imaging techniques present different error characteristics. Therefore, the disadvantages of each of the techniques can be theoretically mitigated by combining the point clouds extracted by each of the techniques to obtain a denser and more accurate representation of the environment.



(a) Passive (b) Active

Figure 3.1: MARESye active and passive data aquisition [2].

Collecting underwater data is a very expensive process to carry out and, for this reason, only a few collections of underwater data or underwater datasets are found in the state of the art. For the same reason, these datasets are generally small, making it difficult to use them to train deep learning algorithms to be used in an underwater environment [57, 58]. Due to the limited availability of relevant underwater datasets, a large-scale synthetic dataset was generated to emulate the data obtained by the MARESye sensor [2]. The large amount of generated data facilitates its use in

training deep learning networks. Since this synthetic data emulates MARESye's output, it enables the trained networks to be effectively applied to real data captured by the sensor.

The dataset employed was generated through a synthetic-to-real approach using Gazebo[2] to simulate the MARESye sensor. This approach enables the creation of 3D data that closely resemble real underwater 3D information. MARESye's simulated sensor configurations and relative orientations comply with the original calibrations. The LSR information is replicated using a 2D laser that simulates a sparse, narrow line of 2.5D points. Given the high reliability and measurement accuracy of LiDAR systems, the dense point cloud was captured using a 128-beam LiDAR. While the LiDAR captures a highly precise and dense group of points from the environment, this information does not match the real output provided by PS stereo. Therefore, the number of points was adjusted to align with the resolution of the actual PS stereo. Additionally, the data collected in simulation were tuned to mirror the noise characteristics and depth resolution of the PS point cloud. The AttentDeepUW network is trained directly with 2.5D projections of 3D point clouds that exclusively represent the depth readings from the collected tridimensional information. Consequently, the absence of texture in synthetic data, compared to real stereo-based information, does not pose a problem. The PS image texture is utilized solely to colorize the final output prediction in real underwater applications. The point clouds are projected onto a 2D plane in the form of depth maps, where each pixel value represents the distance from the camera to a point in the scene. Each 3D point in the point cloud is then multiplied by camera matrix to obtain its 2D coordinates on the image plane. The depth value for each pixel is derived from the Z-coordinate of the transformed 3D point, which represents the distance from the camera to the point.

The stereo input is modeled with characteristics similar to harsh underwater environments [2]:

- The 2D projection of the stereo point cloud is divided into 20 bins, each with a resolution of 0.05 m.

- Adding a random offset within the range of $[-0.05, 0.05]$ m to the depth readings of the entire point cloud is a crucial step in simulating real-world conditions that can impact the accuracy of stereo vision systems. This modification addresses potential issues that could arise from several factors like, poor parameters tuning of the SGBM algorithm, inaccurate stereo calibrations or even mechanical vibrations during operation. By introducing a random depth offset, the dataset more closely mimics these potential inaccuracies, thus providing a more challenging and realistic training scenario for the neural network. This helps ensure that the system is not only effective in ideal conditions but is also robust enough to handle the unpredictable nature of real-world underwater environments.

- Gaussian noise was added to each 3D point, generating multiple noise masks with different means and standard deviations, as detailed in table 3.1. This approach simulates the presence of floating particles, bubbles, and sudden illumination changes typical of underwater

[2] https://gazebosim.org/home

Table 3.1: Parameters of gaussian noise added to dataset.

μ **(mean)**	σ **(standard deviation)**
-0.25	0
0	0.015
0.25	0.03

scenarios, introducing constant discrepancies into the pixel matching operation. By incorporating this erroneous information, the model is better equipped to handle the inherent noise and variability of real underwater environments.

- To effectively mimic real-world radial distortion, which disproportionately affects pixels farther from the center of the image due to camera lens characteristics, noise values were scaled based on each pixel's distance from the center. This was achieved using an inverse distance transform. Pixels received values between 0 and 1 according to their radial distance from the center, with higher values indicating greater distance. Noise masks were then multiplied by these distance values, resulting in a higher concentration of inaccuracies at the image periphery, thus simulating the radial distortion typically observed in camera lenses, as described in [59]. This simulation helps enhance the robustness of the system by preparing it to handle the spatial variability of noise in real underwater imaging scenarios.

Several objects from the ModelNet40 dataset [60] were used as inspiration due to their relevance. The simulated version of MARESye captured 3D information, with 32,000 instances allocated for training and 8,000 for validation. Figure 3.2 presents samples of the data used to train the model. Vertically arranged, the figure includes stereo input, LSR input, and ground truth for each sample. This layout visually illustrates the types of input data provided to the model and the corresponding ground truth used for training, highlighting the variety and structure of the dataset.

For the synthetic dataset in this work, the data was split such that 80% was allocated for training, while the remaining 20% was used for validation. This division ensures that there is sufficient data for both training the model effectively and evaluating its initial performance without overfitting. The data splitting process was implemented using a random selection of point clouds to ensure a diverse representation of scenarios in both the training and validation sets. This approach helps the model to generalize better by exposing it to a wide variety of situations during training.

The model was then tested using real data captured in two settings: a controlled underwater environment (specifically, a clean water tank) and a real maritime environment. This dual testing approach provides a comprehensive evaluation of the model's performance, demonstrating its effectiveness in both ideal and practical conditions.

3.1.1 Data Augmentation

Data augmentation is a widely employed technique to enhance both the volume and variety of data available for training neural networks without collecting new data. This approach involves al-

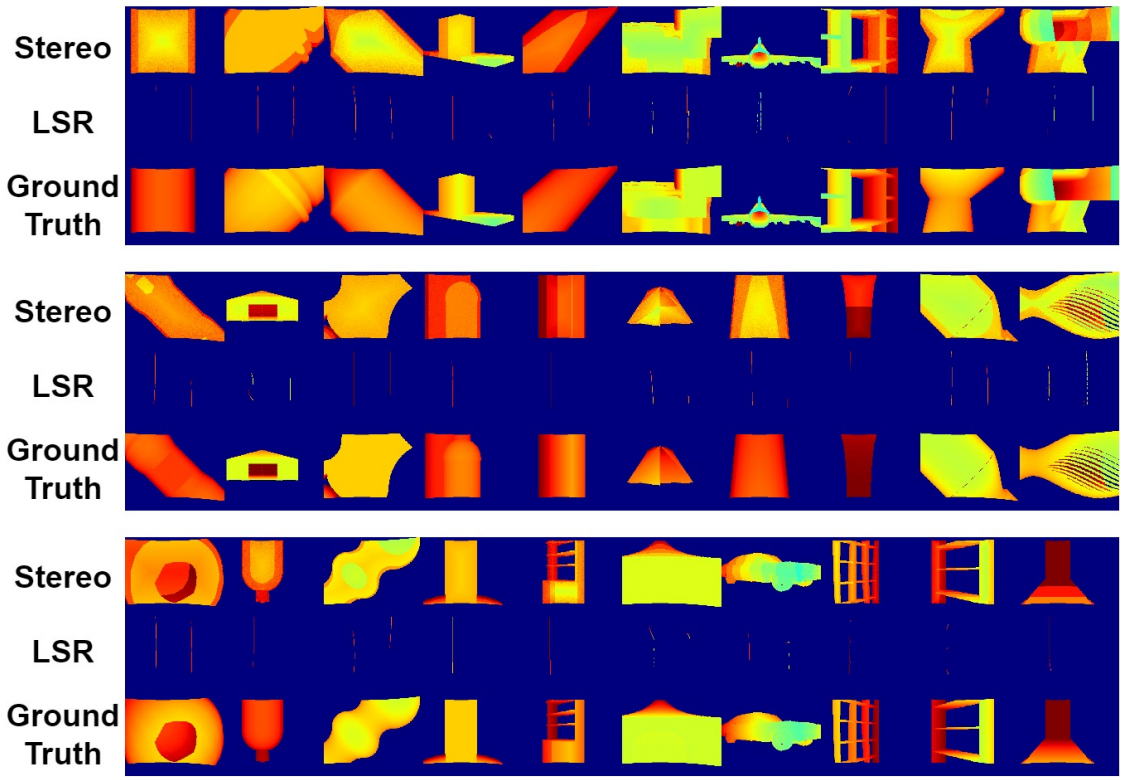


Figure 3.2: Samples of synthetic dataset.

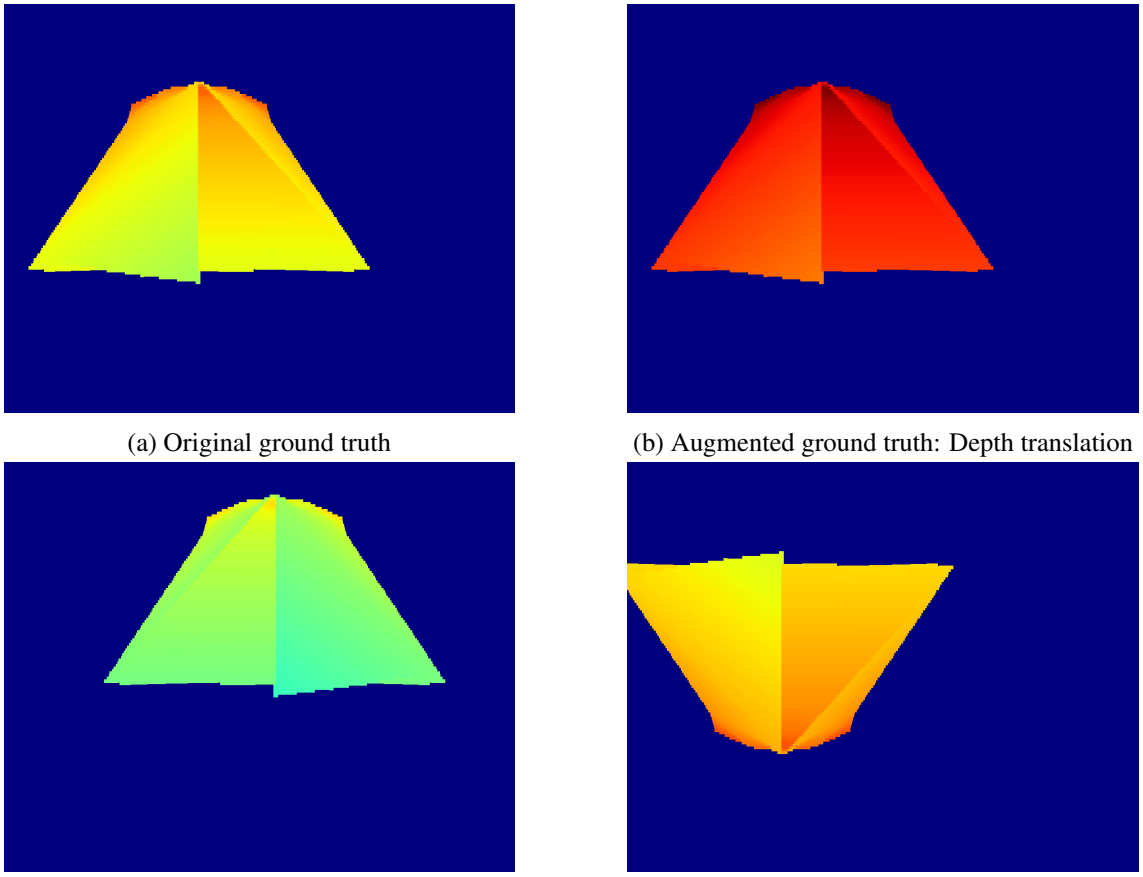
tering each data instance prior to its fusion by the neural network helping to prevent overfitting and make the model more robust to variations in input data. Common approachs for data augmentation includes operations like rotation and cropping, which help to diversify the training examples.
For the specific dataset, various data augmentation techniques were applied to increase the robustness and improve the generalization capability of the model, see table 3.2. The parameters utilized for data augmentation in this study are detailed in table 3.2. This table outlines the specific techniques and corresponding values applied to enhance the diversity and robustness of the dataset. These techniques were systematically selected to ensure effective training across varied scenarios.

These augmentations produces enhance the model's ability to generalize across different scenarios and viewing angles. Depth translations allow the network has a good performance at different depths from very close distances (0.3 m) to far distances (2 m). Projection translations in the data augmentation process enable the network to learn how to reconstruct images independently of the specific region of the image being analyzed. This technique helps improve the network's

Table 3.2: Parameters of Data Augmentations.

Augmentation Technique	**Parameter Specification**
Horizontal Flip	50% of probability of occurrence
Vertical Flip	50% of probability of occurrence
Projection Translation	50% of probability of occurrence in a range of 150 pixels in each direction
Depth Translation	Constant occurrence in a range of -0.25 m to +0.25 m

robustness and ensures consistent performance across different parts of the image. Horizontal and vertical flips are used to further increase the volume of data available for training. By applying these flips with a probability of 50%, there is an equal utilization of both original and augmented data. This balanced approach ensures that the model is exposed to a diverse set of scenarios, enhancing its ability to generalize across different image orientations.



(a) Original ground truth

(b) Augmented ground truth: Depth translation

(c) Augmented ground truth: Depth translation, projection translation and horizontal flip

(d) Augmented ground truth: Depth translation, projection translation and vertical flip

Figure 3.3: Ground truth of a sample and 3 possible augmentations. Are visible the horizontal and vertical flips, the projection translations and different depth translations.

As illustrated in figure 3.3, data augmentation techniques enable the generation of a diverse set of data from a single sample. This approach effectively increases the variety of training examples available, which helps in developing a model that is robust and performs well under different conditions.

3.2 AttentDeepUW Network

The network receives input through stereo and LSR point clouds. After projecting them to 2D, they are concatenated *a priori* to form a combined input. This concatenated input is then processed together as it moves through the encoder, allowing for the simultaneous extraction of features from

both inputs. This approach ensures that the network effectively leverages the integrated data for enhanced feature extraction.

The architecture follows U-Net model [56], which employs an encoder-decoder structure, beginning with an encoder where feature extraction occurs progressively along the network. As the encoder processes the input, it systematically reduces the dimensions of the feature maps. This reduction is designed to streamline the processing workflow and facilitate the extraction of increasingly complex features from the input data. In the decoder segment of the network, the prediction is reconstructed starting with higher-level feature maps.

This process includes skip connections that integrate features from the higher levels of the encoder. This design aids in reconstructing the environment by utilizing more abstract information from the higher levels of the network, which is then combined with more detailed and realistic information provided by the skip connections. The network achieves a more effective reconstruction by concatenation of both detailed and abstract information, allowing it to leverage each source of feature maps optimally. To enhance the flow of information from the encoder to the decoder layers, attention blocks have been incorporated into the network. These blocks enhance the utility of the encoder feature maps in the decoder, thereby improving the quality of the reconstruction by selectively emphasizing critical features [61].

The presented deep neural network is illustrated in figure 3.4, which depicts each layer of the network in detail. This figure provides a comprehensive view of the network architecture, showcasing the arrangement and connection of the layers that constitute the model.

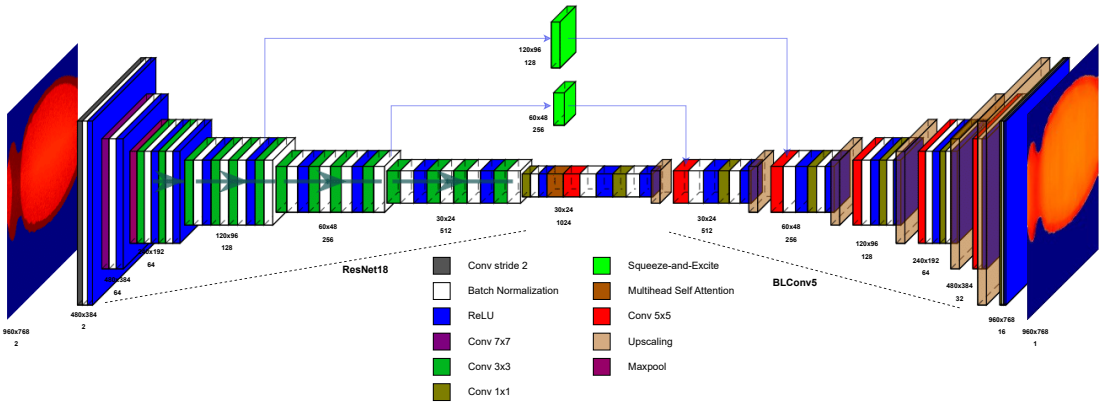


Figure 3.4: Architecture of Network AttentDeepUW.

Encoder

The network initially reduces the dimensions of the input images. This approach allows the network to process smaller feature maps, enabling the use of larger batch sizes. The utilization of larger batches contributes to improved generalization capabilities within the model. This modification involved adding a convolutional block, which consists of a convolution layer with a stride of 2, followed by batch normalization and a ReLU activation layer. This block effectively reduces the

spatial resolution of the feature maps by 75%, reducing them to 25% of their original size. Utilizing a convolutional layer for downsampling facilitates the initial extraction of critical features from the input data, minimizing the loss of important information that might occur with straightforward resizing, where significant details could be lost. The encoder utilizes the ResNet18 architecture, as seen in figure 3.5 which has been pre-trained to enhance its effectiveness in feature extraction and training process. Residual connections in the ResNet18 [62] model address the issue of vanishing gradients by facilitating a more efficient gradient flow. This configuration enables easier network optimization, as it allows gradients to bypass certain layers directly, thereby maintaining their strength throughout the training process. The small complexity of the encoder offers a significant advantage by enabling low inference times, thereby facilitating edge computing. Consequently, inference processes can be directly conducted by the MARESye sensor, streamlining data processing and reducing dependency on external computational resources.

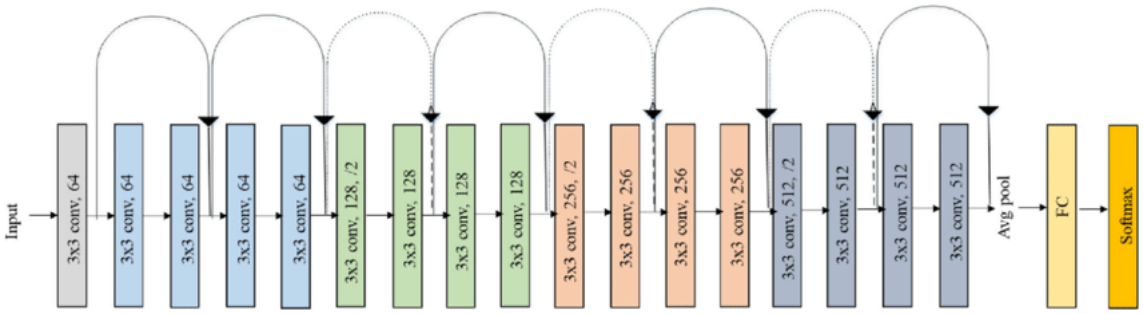


Figure 3.5: ResNet18 Architeture [62].

The standard ResNet18 network receives an input of 3 channels representing an RGB image. However, for the proposed task, the network needs to accept 2 depth maps as input. Therefore, the input convolutional layer of the network must be modified to accommodate the different number of input channels. The first convolution layer was replaced with a similar one in all aspects except for the number of input channels, which was changed from 3 to 2 to accept the projections from the stereo and LSR inputs, respectively. This modification ensures proper adjustment to the number of input channels and allows for appropriate processing of these data. Only the convolutional part of the ResNet18 architecture where features are extracted is utilized in this setup, with the fully-connected layers, typically responsible for classification, being discarded. Additionally, a new convolutional block has been added at the end of the ResNet18 convolutions maintaining the dimensions of the feature maps, further enhancing feature extraction capabilities. The pretraining of ResNet18 was utilized. Nevertheless, the pretraining provides valuable benefits by leveraging learned feature extraction capabilities, enabling the network to handle high-level abstractions more effectively.

Decoder

During the decoding phase, the upsampling task is executed through the BLConv5 decoder [63], which offers a favorable balance between performance and inference times. This decoder

efficiently reconstructs higher-resolution output from the condensed feature maps, ensuring effective image reconstruction while maintaining manageable computational demands. The decoder is structured with five consecutive layers that perform two critical functions: they incrementally increase the spatial dimensions of the output and simultaneously merge the feature maps channel-wise.

Each layer in the decoder is composed of an upsampling step followed by two blocks [63]: depthwise and pointwise. These blocks are composed of a convolution (depthwise convolution and pointwise convolution respectively) followed by a batch normalization and a ReLU layer. The upsampling process within the decoder utilizes the bilinear technique, which is designed to provide a smoother gradient transition between pixels. This method enhances feature maps quality by interpolating the pixel values, resulting in less pixelation during the upscaling process. The bilinear upsampling technique is particularly effective in retaining the clarity and continuity of image features as they are expanded to higher resolutions. The depthwise convolution [64] uses a 5x5 filter to process each output channel independently based on a corresponding input channel. This step allows for efficient spatial filtering as it handles each channel separately. Following the depthwise convolution, the pointwise layer takes over. This layer consists of a 1x1 convolution that consolidates the information from all the input channels into fewer output channels. The pointwise convolution effectively merges the spatially filtered information across all channels, enabling the combination of features extracted by the depthwise layer. This structure enhances the feature integration while maintaining computational efficiency, crucial for effective upsampling in the decoding phase.

For enhanced clarity and understanding, figure 3.6 illustrates the depthwise and pointwise convolutions within a single layer of the decoder architecture. This graphical representation helps elucidate the sequential processing and integration of these convolution types, which are pivotal for refining the feature maps in terms of spatial dimensions and channel-wise information.

This design allows the decoder to effectively reconstruct the original dimensions of the input while integrating diverse feature information across different levels of the network.

Attention Blocks

Skip connections serve as a critical feature in many deep learning architectures, allowing information to bypass certain layers in the network and be directly transmitted from earlier to later layers. This mechanism helps preserve important features and gradients, facilitating more effective learning and deeper network training without the risk of gradient vanishing. However, while skip connections can enhance learning efficiency and model depth, they also have the potential to propagate initial error patterns throughout the network. If the early layers generate inaccurate or misleading feature representations, these errors can be carried forward directly to the output, impacting the overall accuracy and performance of the model [65, 56].

The network proposed in this dissertation has skip connections that connect the outputs of ResNet18 blocks to the inputs of BLConv5 blocks. However, not all blocks are connected via

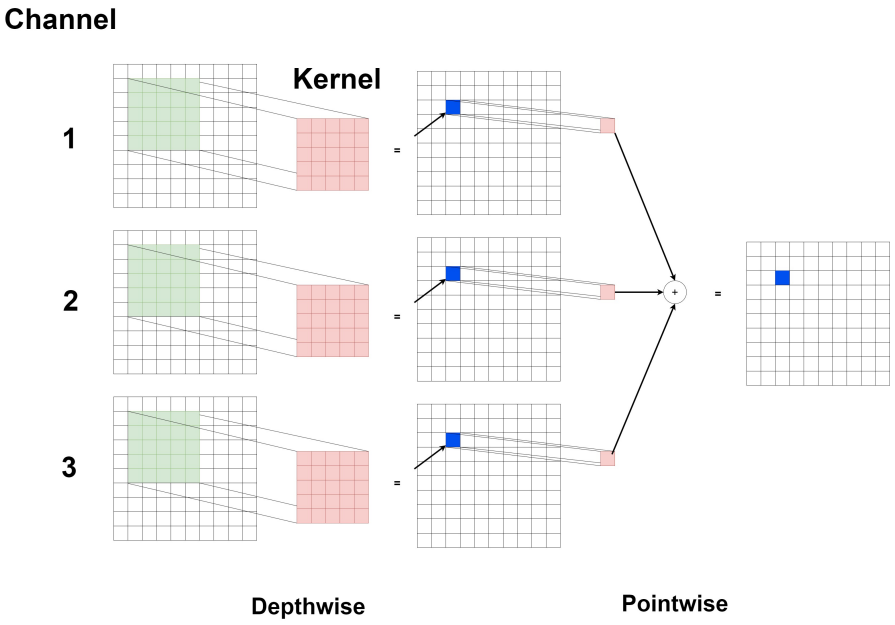


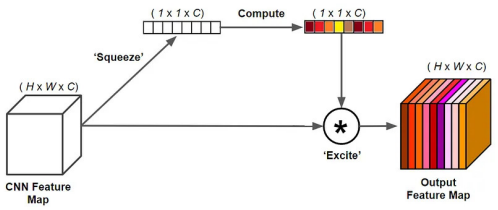
Figure 3.6: Depthwise and Pointwise Convolution.

skip connections. Only the two highest-level skip-connections were added, while the two lowest-level skip-connections weren't implemented to prevent the propagation of errors from the inputs to the output. In the added skip connections, squeeze-and-excitation attention mechanisms [66] were implemented to enhance the network's focus on the most significant channels of the feature maps that are skip connected. The network AttentDeepUW_3skips represents an initial version of the proposed network. The primary distinction between this network and the proposed one is that AttentDeepUW_3skips includes three skip connections, one more than the proposed network. This network was initially proposed before AttentDeepUW but was abandoned during development due to poor performance with real data. This topic is discussed in greater detail in chapter 4. This network only ignores the lowest-level skip-connection.

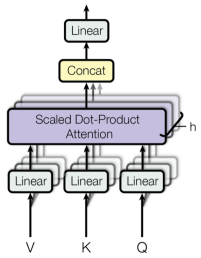
The squeeze-and-excitation mechanism, as shown in figure 3.7a, enhances the model's ability to focus on the most relevant features by recalibrating the channel-wise feature responses. This mechanism operates in two main stages: the "squeeze" stage, where global spatial information is aggregated into a channel descriptor, and the "excitation" stage, where a self-gating mechanism is applied to scale the channel descriptors adaptively. By incorporating this process, the network can prioritize informative features while suppressing less useful ones, leading to improved performance and more efficient learning.

At the network's bottleneck, a Multihead Self Attention mechanism [67], as illustrated in figure 3.7b, adapted to the CNN architecture, has been implemented. This mechanism enables a global context understanding, which is highly advantageous when combined with the local feature extraction capabilities of the standard CNN architecture, creating an effective hybrid approach. The attention layer within this setup is capable of focus its attention where each "head" in the multihead arrangement can attend to different segments of the input data. This allows the model

to capture a diverse array of information from various representational subspaces. While standard CNNs typically excel in extracting spatial hierarchies, they may fall short in capturing contextual relationships within data that are spatially distant. By integrating the Multihead Self Attention mechanism at the network's bottleneck, the model addresses these limitations by facilitating the integration of context over long ranges, thereby enhancing overall model performance. Implementing this attention block at the bottleneck allows the mechanism to access the most critical information and makes the implementation more efficient by processing less data due to the lower resolution of the feature maps at this stage. By prioritizing more important features and de-emphasizing less significant ones, the network becomes more adept at achieving better reconstruction quality and more effective learning outcomes. This focus enhances the model's ability to discern and amplify relevant information, leading to improved performance in tasks such as image reconstruction.



(a) Architecture of Squeeze-and-Excite Block [66].



(b) Architecture of Multihead Attention [67].

Figure 3.7: Architectures of Squeeze-and-Excite Block [66] and Multihead Attention[67].

3.2.1 Network Optimization

The prediction is compared to the associated ground truth, and the discrepancy between them is quantified using a loss function. Based on this error, the network's parameters, specifically the weights, are adjusted by the optimizer to minimize the loss and improve the model's predictive accuracy in subsequent iterations.

The optimization of the network parameters employed a custom l_2-based loss function, as delineated in equation 3.1. This function comprises three components: valid pixels, fill pixels, and neighbor smoothness, which are detailed in equations 3.2, 3.3, and 3.4 respectively.

$$l_2\_ComposedSmothness = 0.75 \times (l_2\_valid) + 0.05 \times l_2\_fill + 0.20 \times l_2\_smothness \tag{3.1}$$

$$l_2\_valid = (y_{px} - \hat{y}_{px})^2, \forall\, y_{px} \in [1e^{-5}, 2.0]\, m \tag{3.2}$$

$$l_2\_fill = (y_{px} - \hat{y}_{px})^2, \forall\, y_{px} \notin [1e^{-5}, 2.0]\, m \tag{3.3}$$

$$l_2\_smothness = (AvgPool_{2D}(y_{px}) - AvgPool_{2D}(\hat{y}_{px}))^2, \forall\, y_{px} \notin [1e^{-5}, 2.0]\, m \tag{3.4}$$

where y_{px} is the ground truth and $\hat{y}_{px}$ is the estimated prediction. The $AvgPool_2D$ operation [3] is described by equation 3.5 with kernel size $k = (15, 15)$, $stride = (1, 1)$ and $padding = 0$. Here, N represents the batch size, C represents the number of channels, h represents the height, and w represents the width of the images.

$$AvgPool_{2D}(N_i, C_i, h, w) = \frac{1}{k[0] \times k[1]} \sum_{m=0}^{k[0]-1} \sum_{n=0}^{k[1]-1} input(N_i, C_j, stride[0] * h + m, stride[1] * w + n) \tag{3.5}$$

The points from the input 3D information, once projected into 2D, are represented as valid pixels. The network will heavily penalize these pixels because they have a direct correspondence to the accurate ground truth. Consequently, these pixels will contribute to 75% of the loss function. On the other side, there are pixels that merely serve to fill the 2D projection in areas where no data exists. These pixels do not contain useful information. To guide the network in learning that these areas are merely fillers, a small weight, constituting 5% of the loss function, is sufficient. The introduction of the $l_2\_smothness$ term, which accounts for 20% of the overall loss function, serves to encourage consistency across neighboring pixels. This aspect of the loss function is designed to ensure that the network doesn't focus only on individual pixel accuracy in depth prediction, but also considers the coherence and continuity among adjacent pixels. This term is crucial for mitigating the impact of noise and artifacts, which are common in sensor data, especially in challenging environments like underwater scenes. By promoting spatial coherence, the model better interprets the essential structures and surfaces, thus making the depth prediction not only accurate but also visually plausible and consistent across the entire image. This approach effectively balances the detailed accuracy needed for individual pixel depth with the broader requirement for a coherent depth perception across the entire field of view. The minimal emphasis on l_2_fill grants the network the flexibility to generate information in regions proximate to valid pixels. This reduced influence of l_2_fill enables pixels, initially assigned a value of zero, to acquire useful information, thereby facilitating the reconstruction of occluded areas using data from neighboring pixels [2].

The learning rate was initially set to 0.1 and strategically reduced by a factor of 1/10 every 3 epochs. This approach is designed to facilitate a precise adjustment of the network's weights as training progresses. Initially, a higher learning rate helps in converging to a good solution quickly, capturing the broad features of the dataset. As the epochs progress, reducing the learning rate helps the model fine-tune its parameters, refining its predictions. This step-wise reduction in the learning rate ensures that the training remains stable and efficient over time, optimizing performance and enhancing the model's ability to generalize from the training data.

[3] https://pytorch.org/docs/stable/generated/torch.nn.AvgPool2d.html#torch.nn.AvgPool2d

Table 3.3: Optimizers used and their updating rules.

Optimizer	Update Rule
SGD [68]	$w_{t+1} = w_t - l_r g_t$
Adam [69]	$w_{t+1} = w_t - \frac{l_r \hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon}$
Adadelta [70]	$w_{t+1} = w_t - \frac{\text{RMS}[\Delta w]_{t-1}}{\text{RMS}[g]_t} g_t$
RMSprop [4]	$w_{t+1} = w_t - \frac{\eta_t}{\sqrt{E[g^2]_t}+\varepsilon} g_t$

Optimizers are algorithms designed to update the parameters of a neural network through back-propagation, aiming to minimize the loss function. The primary goal is to find the optimal values of the weights that yield the most accurate predictions. For training the network, four different optimizers were evaluated: SGD (Stochastic Gradient Descent), Adam, Adadelta, and RMSprop. Each optimizer is characterized by distinct update rules, which dictate how the weights are adjusted during training. Table 3.3 summarizes the update rules for the four selected optimizers:

- **SGD (Stochastic Gradient Descent):** This optimizer updates the weights based on the gradient of the loss function with respect to each weight for each training example, often enhanced with momentum to accelerate convergence.
- **Adam:** Combining the benefits of Adagrad and RMSprop, Adam adjusts learning rates based on the first and second moments of the gradients, providing efficient and reliable convergence.
- **Adadelta:** This optimizer adapts learning rates based on a moving window of gradient updates, making it robust to various hyperparameters.
- **RMSprop:** RMSprop modifies the Adagrad algorithm to reduce its aggressive, monotonically decreasing learning rate, using a moving average of squared gradients to normalize the gradient.

In the given update rules, w_t and w_{t+1} denote the weights at iteration t and $t+1$, respectively. The learning rate is represented by l_r. The gradient of the loss function is denoted as g_t. m_t is the moving average of the gradients, v_t is the squared gradient, and $E[g^2]_t$ is the exponentially decaying average of the squared gradients.

A weight decay parameter of 1×10^{-3} is implemented within the training process to serve as a regularization technique. By doing so, it encourages the network to maintain smaller weight values, making it less likely to fit noise and spurious correlations present in the training data. Regularization via weight decay is crucial for preventing overfitting, particularly in complex neural

[4] https://www.cs.toronto.edu/tijmen/csc321/slides/lecture_slides_lec6.pdf

network models with a large number of parameters. It helps in promoting a simpler model with a better generalization capability on unseen data.

The total number of parameters in the network is 16.7M. This parameter count reflects the complexity of the neural network. Despite the complexity of the network, it proves to be lightweight and capable of being applied to real-time applications. The models generated by training this learning-based approach were obtained using the following hardware: an NVIDIA GeForce RTX 2060 with 6 GB of VRAM and an Intel i5-10600K CPU @ 4.10 GHz with 6 cores. The network was implemented in PyTorch.

Chapter 4

Results and Discussion

This section evaluates the performance of the fusion networks proposed throughout this dissertation. Initially, experiments utilizing synthetic data, as detailed in Section 4.2, illustrate the robustness of the methodologies against noisy inputs. These experiments show a significant enhancement in the quality of the resulting point clouds, both in terms of metrics and visual assessment, when compared to baseline models. Real-world underwater environments introduce complexities that are challenging to replicate in simulations, such as disturbances from floating particles and variable lighting conditions. To bridge the gap between synthetic simulations and real-world applicability, section 4.3 investigates the performance of the synthetic-to-real training approach. This is conducted through precise measurements of the fused 3D outputs within a controlled underwater setting, specifically a clean water tank. The findings from this section aim to validate the effectiveness of the proposed methodologies and their potential for adaptation to more complex and less predictable underwater scenarios.

4.1 Introduction

The testing methodology began by training and validating the AttentDeepUW and AttentDeepUW_3skips architectures, evaluating the performance contributions that each block in the network makes to the final result. After that, tests were then carried out in an underwater environment, characterizing absolute and relative error and then taking relative measurements from a set of objects. Finally, tests were carried out in a real marine environment on the ATLANTIS coastal testbed.

4.2 Synthetic Data Experiments

Overall, the quantitative analysis is guided by the following metrics, which are utilized to evaluate the output predictions:

- **RMSE (Root Mean Squared Error)**

- **MAE (Mean Absolute Error)**
- δ_n**:** The percentage of predicted pixels whose relative error falls within a specified threshold, defined as equation 4.1.

$$\delta_n = \frac{\text{card}\left(\left\{\hat{y} : \max\left(\frac{\hat{y}}{y}, \frac{y}{\hat{y}}\right) < 1.25^n\right\}\right)}{\text{card}\left(\{y\}\right)} \tag{4.1}$$

where y and $\hat{y}$ represent the ground truth and predicted values, respectively, and card denotes the cardinality of a set.

4.2.1 Network Design Evaluation

To assess the performance of the proposed networks, a baseline is established by directly comparing the synthetic stereo input with the ground truth data. This comparison serves as a reference point, allowing for the evaluation of the improvements introduced by the proposed networks. By measuring how closely the synthetic stereo input approximates the ground truth, can be quantified the effectiveness of our approach. This baseline comparison is essential for highlighting the advancements achieved in terms of accuracy and reliability, providing a clear metric for assessing the performance gains attributed to the methodologies introduced in this work.

Table 4.1: Architecture Modifications.

Modification	**RMSE (m)**	**MAE (m)**	δ_1 **(%)**	**RMSE Performance Decline (%)**
AttentDeepUW	0.0167	0.0121	98.9	—
Without convolutional block with stride 2 at the beginning of encoder	0.0253	0.0183	91.3	51.5
Without Multihead Self Attention layer	0.0195	0.0141	98.9	16.8
Without convolutional block at the end of ResNet18 convolutions	0.0186	0.0147	97.9	11.4
Without attention mechanisms added at skip connections	0.0239	0.0195	99.4	43.1
Replace SE blocks with CBAM blocks at skip connections	0.0169	0.0129	99.2	1.2

Without initial dimension redution Without the convolutional block with a stride of 2 that reduces the dimensions of the input projections, the optimization process was limited to using a batch size of 2. This constraint posed significant challenges, particularly in terms of the network's ability to generalize. Small batch sizes can hinder the learning process because they provide less diverse information in each training iteration, which can lead to overfitting and poor generalization to new data. To address the limitations posed by the small batch size and to enhance the network's generalization capacity, was implemented at the beginning of the architecture a convolutional block with a stride of 2. This reduction in size enabled the network to process larger batches of data, thereby improving its generalization capabilities across diverse underwater scenarios.

Unlike a simple resizing operation, which might indiscriminately discard significant details, using a convolutional layer for downsampling allows for an initial extraction of critical features

from the input data. By performing convolution with a stride of 2, the network captures essential information while reducing dimensionality, ensuring that the most relevant features are retained and propagated forward through the network. This method minimizes the loss of important information that might occur with straightforward resizing, where significant details could be lost. To ensure the network's outputs match the original input dimensions for accurate comparison with the ground truth, an upsampling step followed by a convolutional block was incorporated at the end of the network. This upsampling process restores the reduced feature maps to their original resolution, allowing the final predictions to be directly compared with the ground truth data.

Without this initial reduction via the convolutional block with a stride of 2, the proposed network would suffer a performance decrease from 0.0167 m RMSE to 0.0253 m RMSE. This represents a 51.5% performance drop compared to the AttentDeepUW network.

Without Multihead Self Attention layer Subsequently the Multihead Self-Attention (MHSA) mechanism was removed from the network. The MHSA mechanism facilitates a global context understanding, complementing the local feature extraction typically performed by convolutional layers. This enhancement allows the network to capture dependencies across the entire input. To maximize its effectiveness, the MHSA block was added to the bottleneck of the network. The bottleneck is strategically chosen because it contains the highest level and most relevant features extracted by the network. Since MHSA operates on a global context, it is most beneficial when applied to these high-level features, enabling the network to understand and leverage long-range dependencies within the input data. Another advantage of positioning the MHSA block at the bottleneck is the reduced resolution of feature maps at this stage. Lower resolution means fewer data points for the MHSA to process, which translates to lighter computational requirements and faster processing times. This efficiency gain is crucial for maintaining the network's performance while enhancing its capacity for global context comprehension. Without the Multihead Self-Attention (MHSA) mechanism, the proposed network exhibits an RMSE value of 0.0195 m, an MAE of 0.0141 m, and a δ_1 of 98.8%, resulting in a degradation of 16.8% in RMSE performance compared to the AttentDeepUW network.

Without convolutional block extending ResNet18 encoder To further enhance feature extraction, in the AttentDeepUW network, an additional convolutional block was integrated at the end of the ResNet18 encoder, positioned just before the Multihead Self-Attention (MHSA) mechanism. This convolutional block maintains the dimensions of the feature maps while increasing the number of channels, thus enriching the feature representation. This additional convolutional block allows for an extra layer of feature extraction, which in turn improves the quality of features being fed into the MHSA. By increasing the number of channels, the network can capture more complex patterns and nuances in the input data, leading to a more robust and detailed feature set. Positioning this block right before the MHSA ensures that the global context mechanism receives a rich and comprehensive set of features to work with. This strategic placement maximizes the benefits of both local feature extraction and global context understanding, creating a more capable and

nuanced network architecture. Removing this additional convolutional block leads the network to present RMSE metrics of 0.0186 m, MAE of 0.0147 m and δ_1 of 97.9%, increasing the RMSE result by 11.4% comparated to the AttentDeepUW network, thus demonstrating the effectiveness of this convolutional block for the performance of the network.

Without attention mechanisms at skip connections To enhance the network's focus on the most relevant features for scene reconstruction, Squeeze-and-Excite attention blocks were integrated into the skip connections of AttentDeepUW. Attention mechanisms help the network selectively emphasize important features while suppressing less relevant information, thereby improving the overall reconstruction quality. Attention mechanisms are particularly effective in skip connections because they allow the network to dynamically adjust the importance of different features at various stages of processing. By incorporating these blocks, the network can better prioritize critical features that are essential for accurate scene reconstruction.

To evaluate the utility of this attention mechanism, the network's performance was assessed by implementing it with the encoder directly connected to the decoder through simple skip connections. Without the attention blocks, the network produced an RMSE of 0.0239 m, an MAE of 0.0195 m, and a δ_1 of 99.4%. The absence of this attention mechanism reduced the network's capacity by 43.1% in terms of RMSE.

Replace SE blocks with CBAM blocks at skip connections The AttentDeepUW network utilizes Squeeze-and-Excite attention mechanisms in its skip connections. To evaluate the performance of this attention mechanism and determine its suitability, these blocks were replaced with an alternative attention mechanism, the Convolutional Block Attention Module (CBAM). CBAM [71] is an attention mechanism that enhances the network's focus on important features across both spatial and channel dimensions, as illustrated in figure 4.1. This module significantly improves the representational power of Convolutional Neural Networks (CNNs) by directing the network's attention to the most informative regions and channels within the feature maps. The CBAM operates in two sequential stages: channel attention and spatial attention. The channel attention module prioritizes important feature channels, while the spatial attention module focuses on the most relevant spatial regions within the feature maps. This dual attention mechanism allows the network to more effectively capture and emphasize critical aspects of the input data. The application of CBAM blocks to the skip connections resulted in an RMSE result of 0.0169, MAE of 0.0129 and δ_1 of 99.2%, representing 1.2% worse RMSE result compared to the AttentDeepUW network that uses SE blocks.

It is concluded that the application of spatial attention within the skip connections was found to be detrimental to the network's convergence, complicating the training process and ultimately hindering performance. In contrast, channel attention proved beneficial for enhancing scene reconstruction. Consequently, Squeeze-and-Excitation (SE) blocks, which focus exclusively on channel attention, have demonstrated better performance compared to the Convolutional Block Attention

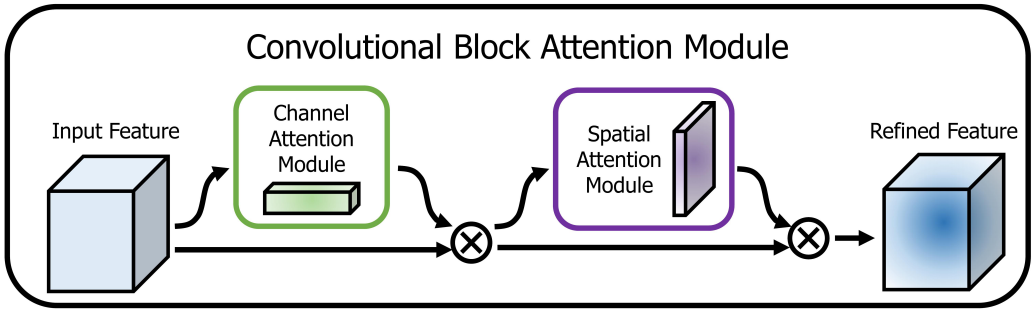


Figure 4.1: Architecture of CBAM [71].

Module (CBAM), which incorporates spatial and channel attention. The focus on channel attention alone proved to be a more advantageous approach, enhancing the network's ability to leverage critical features without the complications introduced by spatial attention. This modification underscores the importance of targeted and efficient attention mechanisms in deep neural network design.

These modifications, as detailed in Table 4.1, underscores the utility of the components that comprise the AttentDeepUW network. The network, AttentDeepUW_3skips, follows the same architectural principles but with an additional low-level skip connection and its respective SE attention block.

4.2.2 Optimizers for training

To effectively train the parameters of the network, several optimizers were tested, including SGD with momentum, Adam, Adadelta, and RMSprop, as detailed in table 4.2. Among these, the SGD with momentum demonstrated superior performance, outperforming the other optimizers.

The use of different optimizers provided valuable insights into the training dynamics of the network. SGD with momentum emerged as the most effective optimizer, offering stable and superior performance. While Adam and RMSprop struggled with convergence issues, Adadelta showed some utility, although with suboptimal results. These findings underscore the importance of optimizer selection in achieving optimal network performance for underwater scene reconstruction.

To determine the best optimization function, three different loss functions were tested: L2 Composed Smoothness with weights of 0.6, 0.05, and 0.35 for valid, fill, and smoothness components respectively; L2 Composed Smoothness with weights of 0.75, 0.05, and 0.2 for valid, fill,

Table 4.2: Optimizers tested for 20 epochs.

Optimizer	**RMSE (m)**	**MAE (m)**	δ_1 **(%)**
SGD	0.0167	0.0121	98.9%
Adam	0.0367	0.0308	51.3%
Adadelta	0.0204	0.0161	98.7%
RMSprop	0.4320	0.4280	52.8%

Table 4.3: Loss functions tested.

Loss	RMSE (m)	MAE (m)	δ_1 (%)
L2 Composed Smoothness 0.6 0.05 0.35	0.0167	0.0121	98.9%
L2 Composed Smoothness 0.75 0.05 0.2	0.0196	0.0142	98.9%
L2 Composed masked	0.0213	0.0145	99.0%

and smoothness components respectively; and L2 Composed Masked, as detailed in table 4.3.

The L2 Composed Masked loss function is an l_2-based loss that distinguishes between valid pixels, which provide useful information, and fill pixels, which merely serve to complete the 2D projection where data is absent. This loss function allocates a contribution of 0.75% to valid pixels and 25% to fill pixels. Among the tested loss functions, the L2 Composed Smoothness emerged as the most effective. This loss function achieved superior results by appropriately weighting valid pixels, which add valuable information, and fill pixels, which do not add useful information and must be learned as merely filling pixels. Additionally, it considers the neighborhood of each pixel, promoting surface consistency. The L2 Composed Smoothness with weights of 0.6, 0.05, and 0.35 yielded better results than the version with weights of 0.75, 0.05, and 0.2. This improvement can be attributed to the higher emphasis on surface consistency in the former configuration. By assigning a greater weight to the smoothness component, the network was better able to ensure consistent surface reconstructions, enhancing the overall quality of the output.

4.2.3 Comparison with state-of-the-art

Table 4.4 presents a comparison of metrics between the proposed fusion methodologies, fusion methodologies previously proposed by P.Leite *et al.* [2] and this baseline. Learning-based methodologies demonstrate greater flexibility in this aspect, effectively improving the input point cloud even with limited information. The network can translate features captured from the sparse input to the surrounding neighborhood pixels. Attention mechanisms aid in selecting the most useful features, thereby enhancing performance in feature utilization. On the downside, despite efforts to enforce neighborhood consistency, the network experiences degradation in this aspect, resulting in artifacts in the predictions and rough gradient patches. Specifically, the edges of the predicted depth maps accumulate the highest amount of error, as illustrated in figure 4.3. These errors can manifest as inaccuracies in depth transitions or inconsistencies in depth values across neighboring pixels.

As demonstrated in figure 4.2, the network AttentDeepUW_3skips produced superior results among all fusion methodologies in synthetic data. However, as detailed in section 4.3, this network has a significant drawback: it propagates input errors to the output prediction. This leads to an overly aggressive correction using LSR information, resulting in data distortion. The network AttentDeepUW, with RMSE metrics of 0.0167 m, 0.0121 m of MAE, δ_1 of 98.8%, despite having worse results than AttentDeepUW_3skips, 0.0147 m RMSE, 0.0108 m MAE, δ_1 of 99.1% on synthetic data, still presents a significant improvement over the baseline, 0.0508 m RMSE, 0.0338 m MAE, δ_1 of 43.5% and RHEA network without preprocessing, 0.0294 m RMSE, 0.0227 m MAE,

δ_1 of 96.7%. Notably, even without preprocessing, AttentDeepUW exhibits comparable metrics to the RHEA network with LSR extended through JMR, 0.0172 m RMSE, 0.0119 m MAE, and δ_1 of 99.0%. The significant discrepancy between AttentDeepUW_3skips and AttentDeepUW is attributed to the capacity of the low-level skip connections to reconstruct the scene. Metrically, AttentDeepUW_3skips exhibits a remarkable improvement of 71.1% compared to the baseline and a notable enhancement of 14.5% relative to the RHEA Network extended through JMR. AttentDeepUW demonstrates a substantial improvement of 67.1% over the baseline and a modest increase of 2.9% compared to the RHEA Network extended through JMR. These metrics underscore the efficacy of both networks in significantly enhancing performance metrics compared to baseline methods.

Table 4.4: Performance comparison between the proposed fusion methodologies with synthetic data, compared to previously proposed methodologies and estimated baseline.

Approach		**RMSE (m)**	**MAE (m)**	δ_1 **(%)**
(a)	Joint Masked Regression (JMR) [2]	0.0428	0.0338	44.5
(b)	RHEA network w/ no preprocessing [2]	0.0294	0.0227	96.7
(c)	RHEA network w/ LSR extended through JMR [2]	0.0172	0.0119	99.0
(d)	AttentDeepUW_3skips w/ no preprocess	**0.0147**	**0.0108**	**99.1**
(e)	AttentDeepUW w/ no preprocess	0.0167	0.0121	98.8
(f)	Baseline	0.0508	0.0476	43.5

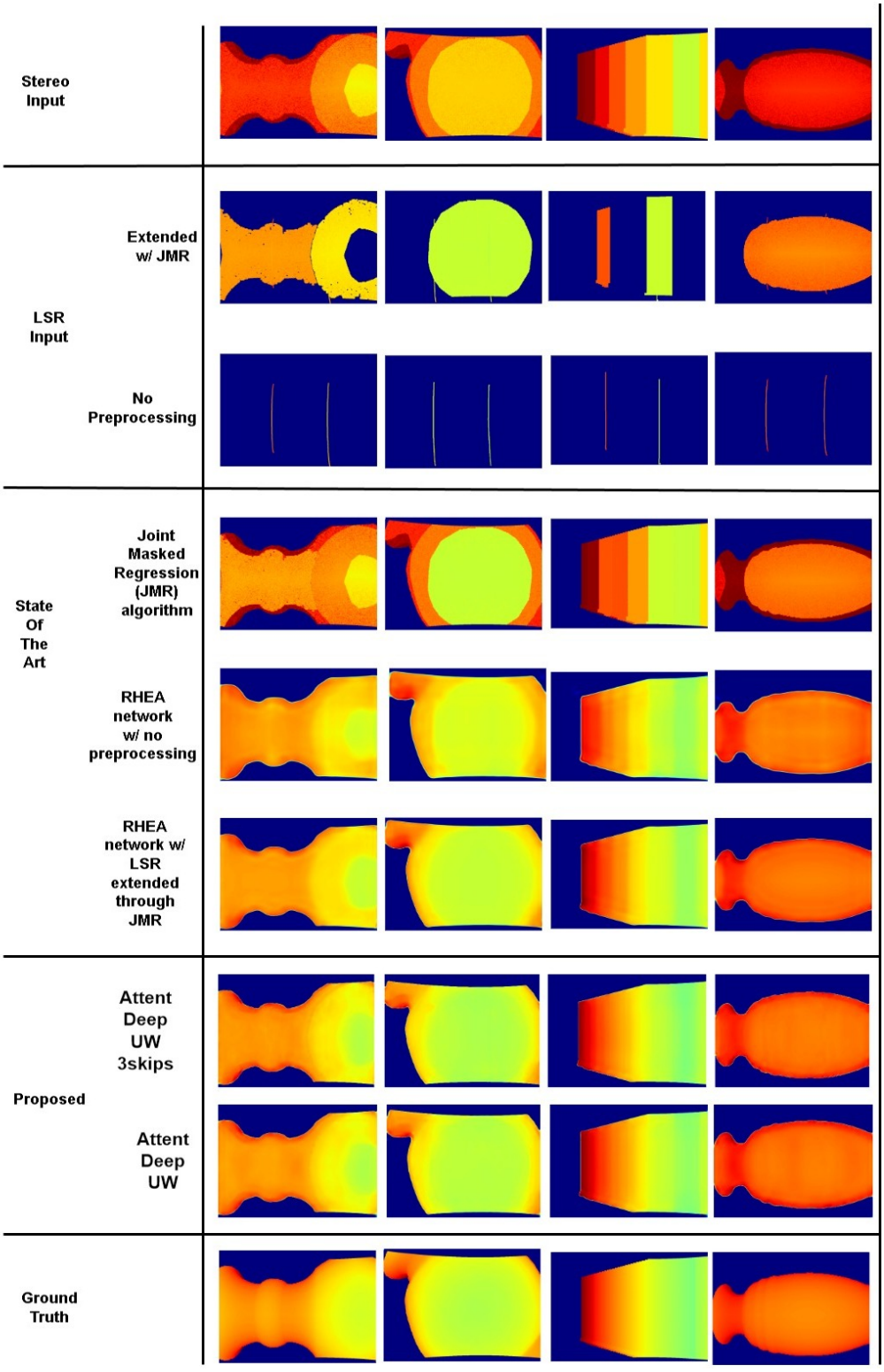
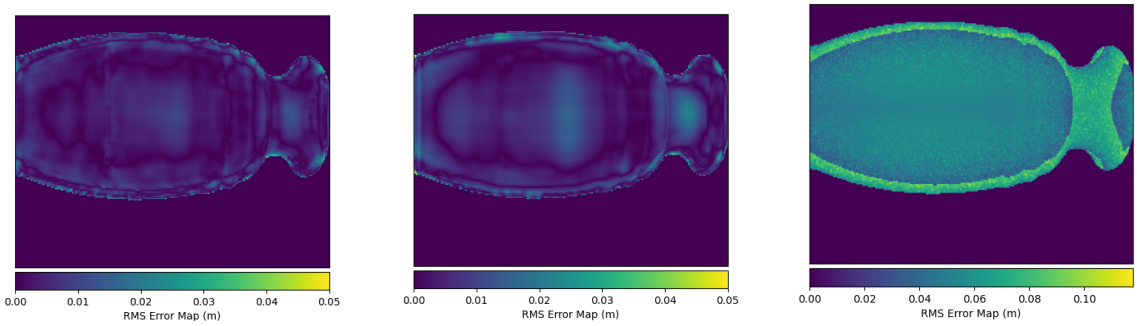


Figure 4.2: Visual comparison of synthetic information between proposed fusion methodologies.



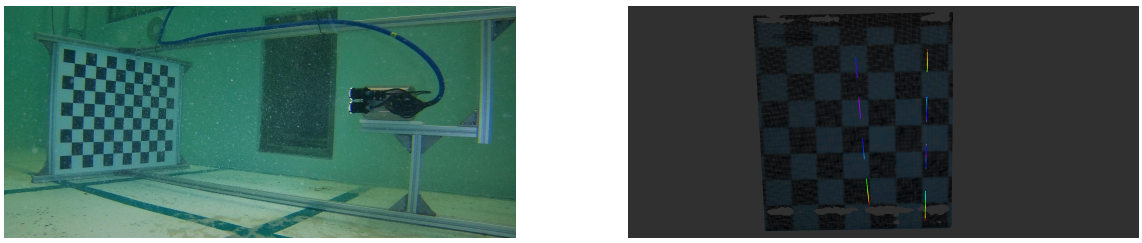
(a) (b) (c)

Figure 4.3: RMSE maps that allow comparing the error in different regions of the projection. RMSE maps of predictions AttentDeepUW_3skips and AttentDeepUW from the last column of figure 4.2 to the ground truth. The baseline error map was calculated by subtracting the stereo input from the ground truth. These maps visualize the root mean square error (RMSE) between the predicted and ground truth depth values, highlighting areas where the predictions deviate from the actual values provided by ground truth data. The predictions AttentDeepUW_3skips and AttentDeepUW correspond to the figures (a) and (b), respectively, while the stereo input corresponds to the figure (c).

4.3 Controlled underwater Experiments

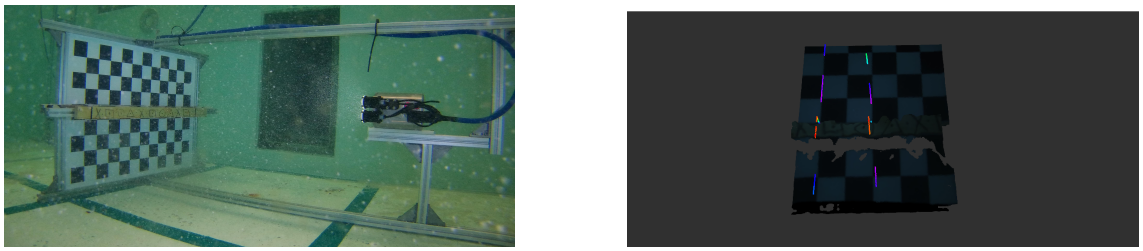
This section delves into the performance evaluation of both the AttentDeepUW and AttentDeepUW_3skips networks, validating the designed synthetic-to-real training methodology through a series of experiments conducted in a controlled environment.

To evaluate the network's performance, an absolute (figure 4.4) and a relative (figure 4.5) error characterization was initially conducted using a chessboard in a controlled underwater environment.



(a) (b)

Figure 4.4: Absolute Error Characterization. Figure (a) shows the experiment to be conducted. Figure (b) exposes the stereo an LSR acquired in absolute error characterization experiment.



(a) (b)

Figure 4.5: Relative Error Characterization. Figure (a) shows the experiment to be conducted. Figure (b) exposes the stereo an LSR acquired in relative error characterization experiment.

For the absolute error characterization, the chessboard was positioned at various distances ranging from 0.35 m to 1.00 m in 0.05 m increments. This setup allowed for a detailed analysis of the network's accuracy at different depths. The relative error characterization followed a similar procedure. A block with dimensions of 0.0902 m in depth and 0.0445 m in height was placed horizontally in front of the chessboard. This chessboard-block combination was then positioned at distances ranging from 0.50 m to 0.85 m, also in 0.05 m increments. This setup provided a comprehensive understanding of the network's performance in distinguishing relative depth differences at various distances. The performance evaluation is conducted by comparing the results obtained with stereo and LSR inputs. This involves a detailed analysis of the output quality, accuracy, and robustness of the 3D reconstruction algorithms.

To evaluate the characterization of absolute and relative error, the distances between the points of twenty-five predictions, for each distance, were measured and averaged to obtain more reliable metrics.

Absolute Error Characterization As depicted in the graph in figure 4.6, both the AttentDeepUW and AttentDeepUW_3skips networks exhibit lower errors compared to the stereo method at distances greater than 0.55 m. In these scenarios, the networks demonstrate an ability to correct the information provided by the stereo input using the LSR data. The AttentDeepUW network demonstrates an average absolute error of 0.0188 m, representing a 65.9% improvement in performance compared to the Photogrammetric Stereo (PS) input point cloud, which has an average absolute error of 0.0551 m. The AttentDeepUW_3skips network shows an average absolute error of 0.0172 m, achieving a 68.8% performance improvement over the PS input point cloud. The improved performance of the AttentDeepUW_3skips network compared to the AttentDeepUW network is attributed to its reduced sensitivity to stereo input. The AttentDeepUW_3skips network assigned significant weight to the LSR input, as evidenced by the graph in Figure 4.6, which closely follows the characteristic error profile of the LSR. Meanwhile, the AttentDeepUW network is influenced more evenly by both the stereo and LSR inputs. It exhibits an error characteristic similar to the stereo input, but with less significant errors, as it utilizes the LSR input to assist in depth correction. However, at a distance of 0.65 m, the AttentDeepUW_3skips network exhibits

Table 4.5: Characterization of Absolute Error Using Chessboard at Various Distances to Assess the Performance of the AttentDeep Network.

Measured Distance (m)	**True Distance (m)**	**Average Chess Distance (m)**	**Average Distance Absolute Error (m)**	**STD Distance Absolute Error (m)**	**Average Relative Error (%)**	**Number of Points**
0,35	0,37604	0,36218	0,01386	0,004379	3,69%	805527
0,4	0,42604	0,40449	0,02155	0,005599	5,06%	917147
0,45	0,47604	0,47408	0,00196	0,007646	0,41%	861371
0,5	0,52604	0,50417	0,02187	0,008763	4,16%	885456
0,55	0,57604	0,56386	0,012184	0,007381	2,12%	865453
0,6	0,62604	0,62400	0,002038	0,008579	0,33%	851729
0,65	0,67604	0,67041	0,005625	0,008703	0,83%	809234
0,7	0,72604	0,73754	0,011500	0,010360	1,58%	818066
0,75	0,77604	0,80019	0,024147	0,012042	3,11%	836908
0,8	0,82604	0,85473	0,028694	0,014886	3,47%	832261
0,85	0,87604	0,91271	0,036669	0,010174	4,19%	836140
0,9	0,92604	0,92494	0,001104	0,009697	0,12%	815024
0,95	0,97604	0,95929	0,016755	0,006961	1,72%	684467
1	1,02604	0,96048	0,065557	0,007300	6,39%	690596

a higher error than the stereo method, indicating a problem and resulting in decreased depth estimation quality. At this distance, the LSR showed an error characteristic of 0.0166 m, slightly higher than expected for this range. Despite the stereo method's reduced error at this distance, with an absolute error of 0.0300 m, the AttentDeepUW_3skips network followed the LSR error characteristic and reached a value of 0.0354 m.

For distances less than 0.55 m, the errors associated with both the stereo and LSR inputs are minimal. In these cases, the AttentDeepUW network shows errors similar to those of the inputs. At distances approaching 1.00 m, the stereo method exhibits significantly elevated errors, which subsequently affect the performance of the network. The results obtained from the absolute error characterization are presented in the tables 4.5 and 4.6 of the AttentDeepUW and AttentDeepUW_3skips networks respectively.

Table 4.6: Characterization of Absolute Error Using Chessboard at Various Distances to Assess the Performance of the AttentDeep_3skips Network.

Measured Distance (m)	**True Distance (m)**	**Average Chess Distance (m)**	**Average Distance Absolute Error (m)**	**STD Distance Absolute Error (m)**	**Average Relative Error (%)**	**Number of Points**
0,35	0,37604	0,35466	0,02138	0,005781	5,69%	860019
0,4	0,42604	0,39523	0,03081	0,004790	7,23%	905652
0,45	0,47604	0,47788	0,00184	0,005972	0,39%	865349
0,5	0,52604	0,50757	0,01847	0,007065	3,51%	862866
0,55	0,57604	0,56741	0,008634	0,007781	1,50%	909349
0,6	0,62604	0,60755	0,018492	0,010343	2,95%	920123
0,65	0,67604	0,64063	0,035411	0,012201	5,24%	919723
0,7	0,72604	0,70670	0,019341	0,017174	2,66%	911396
0,75	0,77604	0,76693	0,009111	0,022536	1,17%	899597
0,8	0,82604	0,81014	0,015898	0,026366	1,92%	892618
0,85	0,87604	0,87191	0,004131	0,027134	0,47%	871129
0,9	0,92604	0,89348	0,032555	0,026339	3,52%	896462
0,95	0,97604	0,96611	0,009929	0,038637	1,02%	889870
1	1,02604	1,01195	0,014088	0,055640	1,37%	865911

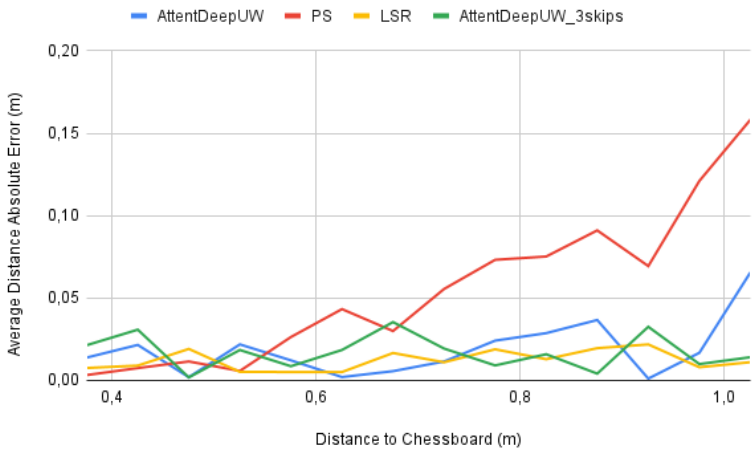
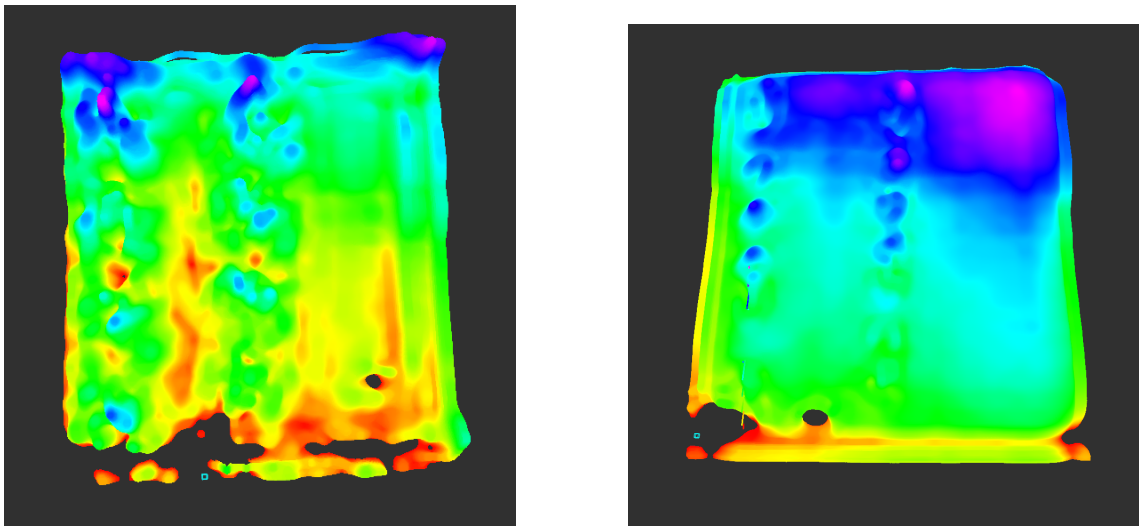


Figure 4.6: Graphic of absolute error characterization.

As illustrated in Figure 4.7, the AttentDeepUW_3skips network tends to produce non-smooth regions around the LSR inputs. This lack of smoothness was a significant factor in the decision to discontinue the use of this network architecture. In contrast, the AttentDeepUW network generates much subtler perturbations, preserving the integrity of the object's shape within the scene. This characteristic makes the AttentDeepUW network more suitable for applications requiring high fidelity in shape reconstruction.



(a) AttentDeepUW_3skips (b) AttentDeepUW

Figure 4.7: Visual comparison between AttentDeepUW_3skips and AttentDeepUW networks in chess images

Relative Error Characterization In relative error characterization, the network's ability to estimate the difference in depth between the chessboard and the block was evaluated. As illustrated in figure 4.8, the AttentDeepUW network consistently demonstrates lower errors compared to both the LSR and stereo methods. However, at distances approaching 0.85 m, the network's errors are influenced by the high error rates associated with the LSR input. In close-range inspections, the short distance to the object allows for significant discontinuities in the laser projections, making it easy to separate segments of the beam that are projected onto and reflected by the chessboard. This enables direct triangulation of the point cloud. However, the same cannot be said for large-range reconstructions. At these distances, in addition to the phenomenon of light absorption, the laser discontinuities become progressively smaller, making it very difficult for the LSR algorithm to estimate depth values for each of the beam's projection planes. Due to the network's reliance on the LSR input primarily for depth estimation, the increase in errors associated with LSR clearly impacted the network's performance. The network demonstrated an average relative error of 0.00616 m, compared to 0.0123 m for the stereo method and 0.0241 m for the LSR. Despite the negative impact of LSR errors at greater distances, the network achieved a 49.9% improvement over the stereo input. Despite this, the AttentDeepUW network still manages to outperform the traditional stereo method across most measured distances, effectively leveraging the LSR data to enhance the accuracy of depth estimation.

The results from the relative error characterization are presented in table 4.7.

Table 4.7: Characterization of Relative Error Using Chessboard at Various Distances to Assess the Performance of the AttentDeep.

Measured Distance (m)	True Distance (m)	Block Depth (m)	STD Error	Absolute Error Block Depth (m)	Average Relative Error (%)	Block Height (m)	Absolute Error Block Height (m)	STD Height	Number of Points Chess	Number of Points Block
0,5	0,52604	0,09421	0,009893	0,00401	4,45%	0,05688	0,012380	0,000406	583469	102615
0,55	0,57604	0,09152	0,011561	0,00132	1,46%	0,06073	0,016229	0,000734	590473	103378
0,6	0,62604	0,09575	0,012047	0,00555	6,15%	0,05824	0,013744	0,001028	748935	80922
0,65	0,67604	0,09812	0,010757	0,00792	8,78%	0,05315	0,008652	0,001508	381338	50413
0,7	0,72604	0,08883	0,015476	0,00137	1,52%	0,05796	0,013459	0,005439	625024	58557
0,75	0,77604	0,08931	0,016782	0,00089	0,98%	0,05459	0,010094	0,003917	544071	53599
0,8	0,82604	0,08135	0,013766	0,00885	9,81%	0,05533	0,010831	0,006879	434038	38432
0,85	0,87604	0,07082	0,012560	0,01938	21,49%	0,06793	0,023432	0,031404	442750	20472

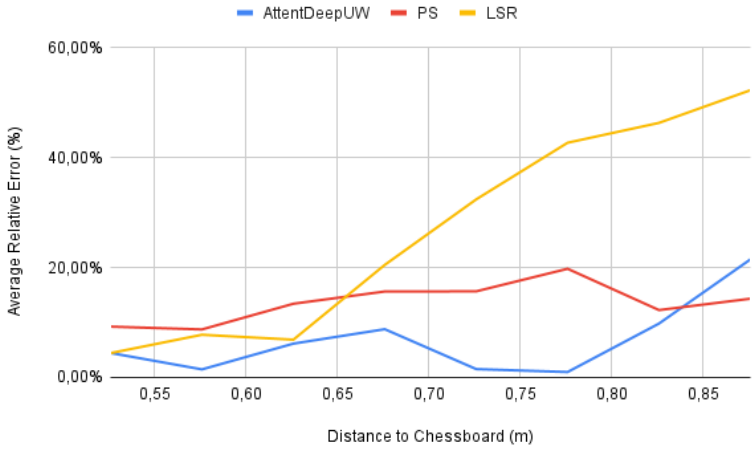


Figure 4.8: Graphic of relative error characterization.

4.3.1 Relative Measurements from a Set of Objects

A predefined set of objects was employed during the trials to mitigate the absence of ground-truth data underwater. Relative measurements of these objects served as the target for metric analysis, as depicted in figure 4.9. In the first experiment, the MARESye system and the objects were positioned in fixed relative positions within a water tank. Point clouds were captured for each object using both Photogrammetric Stereo and Light Stripe Ranging techniques. These point clouds were then fused using the proposed algorithm. The resulting predictions were evaluated based on spatial and volumetric dimensions, as well as distance to the camera. The outcomes of this analysis are summarized in table 4.8.

The AttentDeepUW consistently integrates both input point clouds to achieve a more accurate representation of each object. Consistent with findings from synthetic data, the precise alignment of laser beams is critical for the method's performance. Predictions generated by the AttentDeepUW network demonstrate significant metric improvements over Photogrammetric Stereo

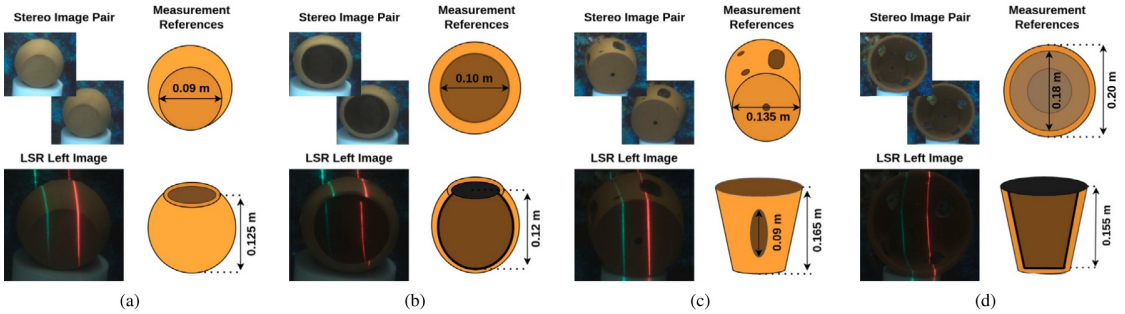


Figure 4.9: The set of objects utilized in the controlled underwater experiments includes various characteristics for analysis. Each object is depicted with pairs of images obtained during the Photogrammetric Stereo mode (enhanced for visualization) and the corresponding reference image acquired during the Light Stripe Ranging stage [2].

estimates. By effectively utilizing all information from the LSR input, the model accurately complements regions of the stereo input. The proposed network demonstrates its capability to compete effectively with the RHEA network in the fusion task. Despite achieving results consistent with other fusion methodologies, the network encountered spatial errors attributed to the initial reduction of network dimensions, which led to the degradation of spatial information. This reduction potentially impacted the network's ability to accurately preserve and utilize spatial details crucial for precise fusion of input data. As a result, spatial errors emerged in the predictions, particularly affecting the fidelity of the reconstructed scenes. Despite the AttentDeepUW network difficulties in predicting spatial dimensions, it proved to outperform stereo and LSR inputs and state-of-the-art methodologies, with an error of 0.0046 m for estimating the inner diameter spatial dimension, and an error of 0.0061 m for the depth volumetric dimension. The high errors in spatial measurements of inner (0.0121 m) and outer (0.0116 m) diameters of D object can be attributed to the network's difficulty in making accurate predictions in areas of high gradient. When the network encounters regions with sharp changes in depth, it tends to smooth out these gradients, resulting in rounded edges instead of the sharp, well-defined boundaries present in the actual objects. This smoothing effect leads to significant inaccuracies in the measurements of critical dimensions, such as inner and outer diameters, because the network fails to preserve the true geometric details of the scene. Consequently, the predicted point clouds exhibit higher measurement errors, particularly in regions where precise delineation of edges is crucial. This issue underscores the challenge of maintaining both smoothness and accuracy in depth estimation, particularly in complex underwater environments where abrupt changes in depth are common.

When considering the volumetric depth of object D, the proposed methodology demonstrates its capability to refine the input point cloud. The AttentDeepUW model excels in estimating volumetric dimensions, achieving a highly precise depth estimation with an error of only 0.0008 m in the output point cloud, as detailed in table 4.8. However, it shows a higher absolute error of 0.0059 m for distance to the camera compared to other methods. The capacity of the proposed network to predict the depth of the scene aligns well with the proposed objectives, surpassing stereo input. Overall, it demonstrates slightly superior performance compared to the RHEA network in terms

Table 4.8: Evaluation of Fusion Methodologies on Underwater Object Dataset despited in figure 4.9. Bold entries denote the highest metric improvement relative to the baseline point cloud.

Object	Characteristics		Ground Truth	Photogrammetric Stereo		Light Stripe Ranging		Joint Masked Regression [2]		RHEA Network [2]		AttentDeepUW Network	
			Measured (m)	Estimated (m)	Absolute Error (m)	Estimated (m)	Absolute Error (m)	Estimated (m)	Absolute Error (m)	Estimated (m)	Absolute Error	Estimated (m)	Absolute Error (m)
A	Spatial Dimensions	Back Diameter	0.0900	0.1030	0.0130	*		0.1030	0.0130	**0.0995**	0.0095	0.1020	0.0120
	Volumetric Dimensions	Height	0.1250	0.0370	0.0880	0.1205	0.0005	0.0370	0.0880	**0.1284**	0.0034	0.1480	0.0230
	Distance to Camera	————	0.4900	0.4968	0.0068	0.4873	0.0027	**0.4932**	0.0032	0.4980	0.0080	0.4700	0.0200
B	Spatial Dimensions	Inner Diameter	0.1000	0.0876	0.0124	*		0.0876	0.0124	0.0924	0.0076	**0.0954**	0.0046
	Volumetric Dimensions	Depth	0.1200	0.1442	0.0242	0.1232	0.0032	0.1270	0.0070	0.1317	0.0117	**0.1261**	0.0061
	Distance to Camera	————	0.5000	0.5049	0.0049	0.5039	0.0039	**0.5016**	0.0016	0.5044	0.0044	0.4870	0.0130
C	Spatial Dimensions	Back Diameter	0.1350	0.1376	0.0026	*		0.1376	0.0026	**0.1328**	0.0022	0.1227	0.0123
	Volumetric Dimensions	Height	0.1650	⋆		**0.1703**	0.0053	⋆		0.1820	0.0170	0.1794	0.0144
	Distance to Camera	————	0.4500	0.4649	0.0149	**0.4469**	0.0031	0.4419	0.0081	0.4422	0.0078	0.4430	0.0070
D	Spatial Dimensions	Inner Diameter	0.1800	**0.1743**	0.0057	*		0.1669	0.0131	0.1692	0.0108	0.1679	0.0121
		Outer Diameter	0.2000	**0.1941**	0.0059			0.1868	0.0132	0.1866	0.0134	0.1884	0.0116
	Volumetric Dimensions	Depth	0.1550	0.1449	0.0101	0.1603	0.0052	0.1583	0.0033	0.1496	0.0054	**0.1558**	0.0008
		Ellipse Carving	0.0090	**0.1010**	0.0110	0.0767	0.0133	0.1141	0.0241	†		†	
	Distance to Camera	————	0.4000	0.4053	0.0053	0.3980	0.0020	**0.3982**	0.0018	0.4035	0.0035	0.3941	0.0059

(*) 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 and AttentDeepUW introduced new points in the point cloud that completely filled the ellipse carving.

of depth estimation. The network's ability to predict depth effectively stems from its architecture, which integrates advanced fusion methodologies and attention mechanisms. By leveraging both stereo and LSR inputs, the AttentDeepUW network harnesses complementary information to produce depth predictions that are more accurate and robust.

However, the flexibility of the convolutional neural network can sometimes introduce unreliable information. Object D in table 4.8 serves as an example: the ellipse carving, measured as 0.1010 m in the stereo input, is completely filled in by the network, interpreting the carving as missing information rather than a feature of object D. Despite efforts in training to enforce neighborhood consistency, artifacts persist in the predictions, especially around high-gradient regions like the edges of the point cloud. These experiments demonstrate the effectiveness of the proposed fusion network in underwater environments. The stereo input often lacks information due to occlusion zones, resulting in less precise and faithful output predictions unless both stereo and LSR inputs overlap, allowing for more accurate predictions. The findings from synthetic data transfer effectively to real-world applications, validating the proposed synthetic-to-real training methodology. The network excels in estimating improved output point clouds by leveraging information from both inputs, particularly guided by the LSR input to fill in missing details. However, while these additions improve results in certain scenarios, they can also introduce erroneous information, especially in high-gradient regions.

4.4 ATLANTIS coastal testbed - real maritime environment

In contrast to controlled underwater environments, real underwater environments are characterized by harsh conditions such as sediment presence and constant light changes. These factors introduce additional challenges including reduced visibility and increased backscattering of photons. To test the robustness of the proposed fusion methodology under these conditions, data was collected at the ATLANTIS Coastal Test Centre, quayside in Viana do Castelo's Port infrastructure [1]. The offshore floating structure DURIUS, repurposed from the Oil & Gas industry, provided an ideal platform for this testing. Access to this platform allowed for the collection of data using the MARESye sensor, and for testing the data fusion algorithms in a real-world application [2].

The sensor was used to gather information on specific areas of interest on the buoy. Multiple data collection campaigns were conducted under diverse environmental conditions, resulting in varying amounts of suspensoids within the water. The heterogeneous tridimensional information retrieved was then fused using the AttentDeepUW network. The results of these experiments are discussed in this section [2].

(a) (b)

Figure 4.10: The data was gathered at the ATLANTIS Coastal Testbed located in Viana do Castelo. In the real application experiments, DURIUS, the floating structure, is depicted in Figure (a). Figure (b) shows how the team at INESC TEC is collecting information with the help of MARESye, which is suspended from a support vessel [2].

The data collected from the DURIUS floating buoy enabled the validation of the proposed fusion algorithm in a real underwater environment. Figure 4.11 illustrates several instances from these trials. The analysis was limited to a qualitative discussion due to the absence of ground-truth data. During the tests, the DURIUS was covered with biofouling, both helping and challenging the collection of 3D information by MARESye. The photogrammetric stereo method benefited from the rich textures of the environment, enhancing its performance. In contrast, this scenario became particularly challenging due to the absorption of the laser beam in sediment heavy waters and the fact that many bio-organisms shared the same color as one of the projected laser beams (green - 520 nm), significantly hindering LSR data collection and making segmentation impossible for Instances 1, 2, and 3. Despite having only a single laser beam as input, the AttentDeepUW network effectively utilized the available sparse information to generate a more accurate point cloud [2].

Instance 1 in figure 4.11 shows the top part of a pillar structure with a thick border of attached mussels, representing a relatively smooth region. The AttentDeepUW network successfully reconstructs the smoothness of the surface. However, the network was unable to perfectly align the output prediction with the laser beam line. The resulting point cloud is positioned at a depth between the laser beam and the stereo input, closer to the laser beam, and potentially nearer to the actual depth information. This misalignment is attributed to the network's limited ability to fuse information at very close depths.

In Instance 1, both the RHEA and AttentDeepUW networks introduce curvature to the point cloud, accurately depicting the top border of the cylindrical structure. However, the RHEA network incorrectly propagates this curvature to the remaining portion of the point cloud, as highlighted in green in Instance 1 of figure 4.11. In contrast, the AttentDeepUW network successfully introduces curvature to the top border of the cylindrical structure without extending this curvature to the rest of the point cloud. On the other hand, the AttentDeepUW network didn't accurately correct the global position of the point cloud relative to the camera, maintaining a distance between the output point cloud and the LSR, as indicated by the red arrows in Figure 4.11.

A continuation of this pillar structure is depicted in Instance 2 of figure 4.11. In this instance, the point clouds are situated slightly farther from the camera, which enhances the network's performance. Here, the AttentDeepUW network not only maintains the smoothness of the environment but also successfully aligns the output prediction with the laser beam. This accurate superposition indicates that the network effectively fuses the stereo and LSR inputs, achieving a more precise depth estimation. The increased distance likely provides the network with more distinguishable depth cues, enabling better fusion and alignment of the point clouds. The AttentDeepUW network demonstrates high robustness when working with limited information, as it was trained on scenarios where only single-beam data was available. Despite this, the AttentDeepUW network enhances the smoothness of the output predictions compared to the RHEA network, while maintaining the quality of depth estimation. This improvement is evident in the output predictions shown in figure 4.11. The AttentDeepUW network successfully regresses the entire input point cloud, even with the sparse laser information, proving its capability to produce more accurate and smoother depth maps under challenging conditions.

Instance 3 of figure 4.11 presents a scenario where one of the beams supporting the stairs of DURIUS is heavily covered with fouling, enhancing the already complex depth gradient. In this instance, the laser beam information is optimally positioned to convey part of the structure's volumetric dimensions. Consequently, both tri-dimensional fusion algorithms excel in this situation, successfully regressing the input point cloud based on the sparse information. Additionally, the AttentDeepUW network introduces points into the output prediction, filling in previously empty sections of the point cloud caused by occlusion zones. The output prediction of AttentDeepUW closely resembles the prediction generated by the RHEA network.

Despite the initial data collection campaign using MARESye at the ATLANTIS Coastal Testbed successfully gathering high-quality information, which validated the 3D fusion capability of the Joint Masked Regression (JMR) algorithm, the RHEA network, and the AttentDeepUW network

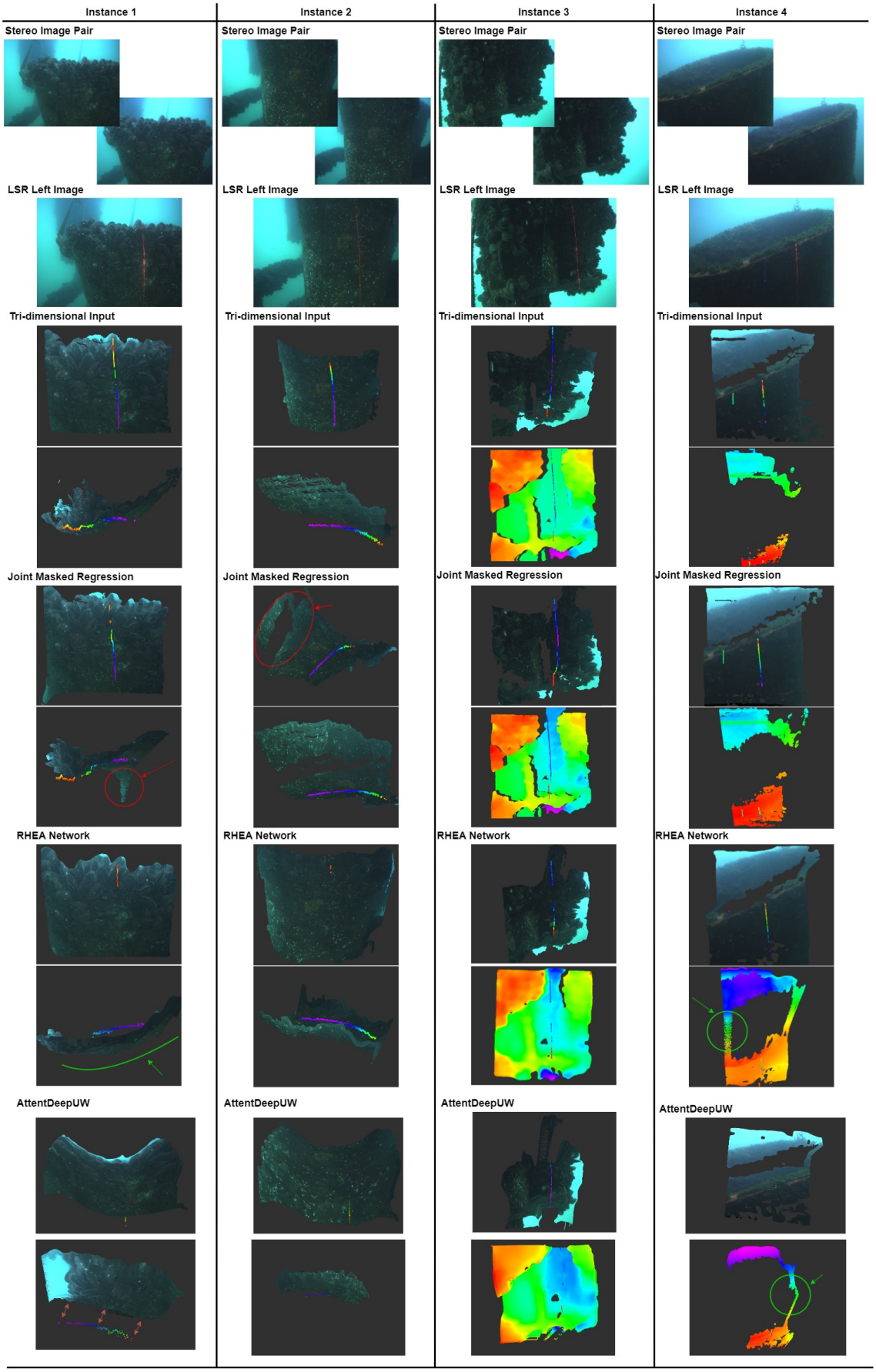


Figure 4.11: Comparison fusion methodologies using real underwater data from the MARESye sensor at the ATLANTIS Coastal Testbed.

in a real underwater environment, the fact that only one laser was functional impacted the visual results. This impact was most pronounced in the performance of the JMR algorithm. Based on these insights, improvements were made to the MARESye imaging system by replacing the green laser beam with a violet one (405 nm) to facilitate laser segmentation and avoid confusion with existing bioorganisms [2]. Subsequent data collection campaigns occurred under challenging conditions, including heavy rain and increased turbidity caused by floating particles, significantly altering the environmental dynamics [2]. Instance 4 in figure 4.11 illustrates one of the experiments conducted under these environmental conditions, focusing on a cylinder-shaped structure similar to that of the DURIUS buoy. In this trial, the AttentDeepUW network performs as anticipated, generating a point cloud that aligns with the LSR input and exhibits increased density due to the network introducing additional points. The network effectively corrects the input point cloud while preserving its curved shape, as demonstrated in the lower part of instance 4 in figure 4.11. The presence of two available laser beams contributes to this improved outcome, as the regression curve does not need to generalize from a single beam across the entire uniform region — a limitation observed in Instances 1 and 2. The green-circled points added on the right side contribute to completing the structure's shape. Compared to RHEA, AttentDeepUW demonstrates improvement by avoiding the erroneous connection of both portions of the cylinder.

Chapter 5

Conclusions and Future Work

The integration of various sensors and the development of advanced perception systems are crucial for improving offshore wind energy operation and maintenance through enhanced underwater perception. By leveraging heterogeneous data fusion and addressing the challenges of underwater perception, robotic systems can navigate, interact, and operate efficiently in complex and dynamic underwater environments, ultimately reducing costs and improving the reliability of offshore wind energy production. The heterogeneous 3D information captured by the MARESye sensor, which includes a dense and textured Photogrammetric Stereo (PS) point cloud along with multiple sparse yet highly accurate lines of points triangulated via Light Stripe Ranging (LSR), can be effectively combined into a single, dense, and precise representation. The proposed AttentDeepUW method is a learning-based approach that utilizes early fusion to jointly learn features from a coupled representation of both tri-dimensional inputs. By employing attention mechanisms, the network's learning and overall performance are significantly enhanced.

The network's optimization leverages a synthetic-to-real training scheme, effectively bypassing the need for domain-adaptation methodologies. This approach facilitates the direct deployment of the network in underwater scenarios, ensuring robust and reliable performance in real-world applications. Extensive experiments have been conducted to validate the proposed fusion algorithms across various settings, including simulation, controlled environments, and real-world applications with data collected from the DURIUS platform. These comprehensive tests demonstrate the robustness and effectiveness of the developed methodologies. The synthetic data experiments results in metrics RMSE of 0.0167 m, 0.0121 m MAE, and δ_1 of 99.1% to AttentDeepUW.

In 3D fusion error characterization, the AttentDeepUW network exhibited an average absolute error of 0.0188 m. This performance signifies a metric improvement of 65.9% compared to the input Photogrammetric Stereo (PS) point cloud. Such a significant enhancement underscores the network's ability to accurately reconstruct the 3D structure of underwater scenes, correcting the inherent inaccuracies present in the PS point cloud data. Furthermore, the network also achieved an average relative error of 0.00616 m. This represents an improvement of 49.9% over the input PS point cloud. This metric reflects the network's proficiency in maintaining the relative spatial relationships within the 3D data, ensuring a more coherent and precise representation of the

underwater environment.

The object experiment in controlled underwater allowed an evaluation of the network's performance on real objects, enabling a spatial and volumetric evaluation of the network's predictions in an underwater environment. This experiment allowed to conclude that the network causes a degradation of spatial dimensions caused by the initial reduction of the network resulting in an average absolute error of 0.0105 m. As for volumetric dimensions, the network was able to produce predictions with an average absolute error of 0.0111 m. The experiments conducted at the ATLANTIS Coastal Testbed introduced significant challenges due to harsh underwater conditions and the presence of biofouling on the DURIUS structure, which impeded the data acquisition process. Despite these challenges, the AttentDeepUW network demonstrated its robustness and effectiveness in producing an improved output point cloud. The AttentDeepUW network successfully predicted enhanced output point clouds even when provided with a sparse set of input features, validating the efficacy of the synthetic-to-real training scheme employed. This approach allows the network to adapt seamlessly to real-world underwater scenarios without requiring additional domain adaptation methodologies. Moreover, the network consistently generated dense point clouds that were accurately adjusted based on LSR (Light Stripe Ranging) information. The resulting point clouds exhibited smooth surfaces, reflecting a high degree of consistency and precision. This capability is particularly notable given the unfavorable conditions where only one of the laser beams could be segmented.

In these challenging environments, the network not only maintained good performance but also demonstrated its capacity to extend the information extracted by the LSR, effectively enhancing the overall quality of the 3D reconstruction. This adaptability and resilience underscore the network's practical utility in real-world underwater applications, where data acquisition is often hindered by environmental factors. Throughout the conducted tests, the network consistently produced accurate output predictions, with an average processing time of approximately 4.2 milliseconds per prediction. This efficiency underscores the network's potential for real-time applications. In conclusion, the proposed fusion network has demonstrated remarkable robustness in handling noisy and harsh environmental conditions. In addition to this improvement, further development can be conducted, such as the following:

- **Enhancing Network Robustness and Adaptability:** future work can focus on refining the network to produce an even more robust and adaptable response to 3D heterogeneous data.
- **Implementing MARESye Sensor and Fusion Algorithms in Robotic Systems:** another promising direction is the development of approaches to integrate the MARESye sensor and fusion algorithms with a robotic arm. This integration would create an eye-in-hand system, enabling precise and dynamic 3D mapping and inspection capabilities in complex underwater environments.

References

[1] 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, June 2023.

[2] Pedro Nuno Leite and Andry Maykol Pinto. Fusing heterogeneous tri-dimensional information for reconstructing submerged structures in harsh sub-sea environments. *Information Fusion*, 103, 2024.

[3] Shitong Hou, Dai Jiao, Bin Dong, Haochen Wang, and Gang Wu. Underwater inspection of bridge substructures using sonar and deep convolutional network. *Advanced Engineering Informatics*, 52:101545, 4 2022.

[4] Rafael Marques Claro, Diogo Brandão Silva, and Andry Maykol Pinto. Artuga: A novel multimodal fiducial marker for aerial robotics. *Robotics and Autonomous Systems*, 163, 2023.

[5] 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, 2021.

[6] 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, 2024.

[7] Pedro Leite, Renato Silva, Aníbal Matos, and Andry Maykol Pinto. An hierarchical architecture for docking autonomous surface vehicles. In *2019 IEEE International conference on autonomous robot systems and competitions (ICARSC)*, pages 1–6. IEEE, 2019.

[8] Maria Inês Pereira, Rafael Marques Claro, Pedro Nuno Leite, and Andry Maykol Pinto. Advancing autonomous surface vehicles: A 3d perception system for the recognition and assessment of docking-based structures. *IEEE Access*, 9, 2021.

[9] 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.

[10] John McConnell, Ivana Collado-Gonzalez, and Brendan Englot. Perception for underwater robots. *Current Robotics Reports*, 3, 2022.

[11] Andry Maykol Pinto and Anibal C. Matos. Maresye: A hybrid imaging system for underwater robotic applications. *Information Fusion*, 55, 2020.

[12] Albert Palomer, Pere Ridao, Dina Youakim, David Ribas, Josep Forest, and Yvan Petillot. 3d laser scanner for underwater manipulation. *Sensors (Switzerland)*, 18, 2018.

[13] 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.

[14] Huimin Lu, Yujie Li, Yudong Zhang, Min Chen, Seiichi Serikawa, and Hyoungseop Kim. Underwater optical image processing: A comprehensive review. *Mobile networks and applications*, 22:1204–1211, 2017.

[15] Silvia Corchs and Raimondo Schettini. Underwater image processing: State of the art of restoration and image enhancement methods. *EURASIP Journal on Advances in Signal Processing 2010 2010:1*, 2010:1–14, 4 2010.

[16] Dongsheng Guo, Yiqing Huang, Tianshun Han, Haiyong Zheng, Zhaorui Gu, and Bing Zheng. Marine snow removal. In *OCEANS 2022-Chennai*, pages 1–7. IEEE, 2022.

[17] Yvan Petillot, Ioseba Tena Ruiz, and David M. Lane. Underwater vehicle obstacle avoidance and path planning using a multi-beam forward looking sonar. *IEEE Journal of Oceanic Engineering*, 26:240–251, 4 2001.

[18] N. Brahim, S. Daniel, and D. Guériot. Potential of underwater sonar systems for port infrastructure inspection. In *OCEANS 2008*, pages 1–7. IEEE, 2008.

[19] Pedro V. Teixeira, Michael Kaess, Franz S. Hover, and John J. Leonard. Underwater inspection using sonar-based volumetric submaps. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 2016-November, pages 4288–4295. IEEE, 2016.

[20] Youngseok Kim and Jaesuk Ryou. A study of sonar image stabilization of unmanned surface vehicle based on motion sensor for inspection of underwater infrastructure. *Remote Sensing*, 12(21):3481, 2020.

[21] Thomas Guerneve and Yvan Petillot. Underwater 3d reconstruction using blueview imaging sonar. In *OCEANS 2015-Genova*, pages 1–7. IEEE, 2015.

[22] 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 2013, Vol. 13, Pages 11007-11031*, 13:11007–11031, 8 2013.

[23] Andrew Orriordan, Thomas Newe, Gerard Dooly, and Daniel Toal. Stereo vision sensing: Review of existing systems. *Proceedings of the International Conference on Sensing Technology, ICST*, 2018-December:178–184, 7 2019.

[24] Matija Rossi, Petar Trslić, Satja Sivčev, James Riordan, Daniel Toal, and Gerard Dooly. Real-time underwater stereofusion. *Sensors 2018, Vol. 18, Page 3936*, 18:3936, 11 2018.

[25] Xiyan Sun, Yingzhou Jiang, Yuanfa Ji, Wentao Fu, Suqing Yan, Qidong Chen, Baoguo Yu, and Xingli Gan. Distance measurement system based on binocular stereo vision. In *IOP Conference Series: Earth and Environmental Science*, volume 252, page 052051. IOP Publishing, 2019.

[26] Achuta Kadambi, Ayush Bhandari, and Ramesh Raskar. 3d depth cameras in vision: Benefits and limitations of the hardware with an emphasis on the first-and second-generation kinect models. *Advances in Computer Vision and Pattern Recognition*, 67:3–26, 2014.

[27] Adrian Kaehler and Gary Bradski. *Learning OpenCV 3: computer vision in C++ with the OpenCV library*. " O'Reilly Media, Inc.", 2016.

[28] Mark R Shortis, A Williams, B A Barker, and M Sherlock. A towed body stereo-video system for deep water benthic habitat surveys. *Eighth Conf. Optical*, pages 150–157, 2007.

[29] Kresimir Williams, Alex De Robertis, Zachary Berkowitz, Chris Rooper, and Rick Towler. An underwater stereo-camera trap. *Methods in Oceanography*, 11:1–12, 12 2014.

[30] 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*, 48:200–205, 1 2015.

[31] Marc Carreras, Juan David Hernandez, Eduard Vidal, Narcis Palomeras, David Ribas, and Pere Ridao. Sparus ii auv - a hovering vehicle for seabed inspection. *IEEE Journal of Oceanic Engineering*, 43:344–355, 4 2018.

[32] 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*, 9 2015.

[33] 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, 2023.

[34] Vadim Kramar, Aleksey Kabanov, Oleg Kramar, Sergey Fateev, and Valerii Karapetian. Detection and recognition of the underwater object with designated features using the technical stereo vision system. *Fluids*, 8, 2023.

[35] 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*, 48(16):113–120, 2015.

[36] Gabrielle Inglis, Clara Smart, Ian Vaughn, and Chris Roman. A pipeline for structured light bathymetric mapping. *IEEE International Conference on Intelligent Robots and Systems*, pages 4425–4432, 2012.

[37] Flavio Lopes, Hugo Silva, Jose Miguel Almeida, Alfredo Martins, and Eduardo Silva. Structured light system for underwater inspection operations. In *OCEANS 2015-Genova*, pages 1–6. IEEE, 2015.

[38] Miquel Massot-Campos and Gabriel Oliver-Codina. Underwater laser-based structured light system for one-shot 3d reconstruction. *Proceedings of IEEE Sensors*, 2014-December:1138–1141, 2014.

[39] Miguel Castillon, Albert Palomer, Josep Forest, and Pere Ridao. Underwater 3d scanner model using a biaxial mems mirror. *IEEE Access*, 9, 2021.

[40] 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, 2023.

[41] 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. *Available at SSRN 4751399*, 2024.

[42] Joaquim Salvi, Sergio Fernandez, Tomislav Pribanic, and Xavier Llado. A state of the art in structured light patterns for surface profilometry. *Pattern Recognition*, 43, 2010.

[43] Jason Geng. Structured-light 3d surface imaging: a tutorial. *Advances in Optics and Photonics*, 3, 2011.

[44] 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, 2011.

[45] Amin Sarafraz and Brian K. Haus. A structured light method for underwater surface reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 2016.

[46] 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 & Laser Technology*, 43:801–805, 6 2011.

[47] 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, 2 2023.

[48] Zhenmin Zhu, Hongwei Qiu, Qiang Hu, Kang Ren, Lisheng Zhou, and Taowei Zhu. Underwater 3d reconstruction based on double n-step orthogonal polarization state phase shift strategy. *Optics and Lasers in Engineering*, 178, 2024.

[49] 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, 2013.

[50] Will Maddern and Paul Newman. Real-time probabilistic fusion of sparse 3d lidar and dense stereo. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 2016-November, pages 2181–2188. IEEE, 2016.

[51] 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. IEEE, 2018.

[52] 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. IEEE, 2018.

[53] 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.

[54] Xuelian Cheng, Yiran Zhong, Yuchao Dai, Pan Ji, and Hongdong Li. Noise-aware unsupervised deep lidar-stereo fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 6339–6348, 2019.

[55] 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, 2024.

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, volume 9351, pages 234–241. Springer, 2015.

[57] 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. IEEE, 2023.

[58] 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, 2020.

[59] 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, 2007.

[60] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:1912–1920, 2015.

[61] 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. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12131 LNCS:379–392, 2020.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 2016-December, pages 770–778, 2016.

[63] Pedro Nuno Leite and Andry Maykol Pinto. Exploiting motion perception in depth estimation through a lightweight convolutional neural network. *IEEE Access*, 9, 2021.

[64] Yunhui Guo, Yandong Li, Liqiang Wang, and Tajana Rosing. Depthwise convolution is all you need for learning multiple visual domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8368–8375, 2019.

[65] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[66] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42, 2020.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 2017-December, 2017.

[68] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 4747, 2016.

[69] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[70] Delona C Johny, Anju J S Assistant, Ruirui Zhang, Xin Xiao, Nasir Rashid, Javaid Iqbal, Fahad Mahmood, Anam Abid, Umar S. Khan, Mohsin I. Tiwana, Zhang Fan, Chen Wen, Li Tao, Cao Xiaochun, Peng Haipeng, Chao Yang, Lin Jia, Bing Qiu Chen, Hai Yang Wen, Matthew D. Zeiler, John E. Hunt, Denise E. Cooke, Stephanie Forrest, Alan S Perelson, Lawrence Allen, and Rajesh Cherukuri. Adadelta: An adaptive learning rate method. *IEEE Access*, 7, 2018.

[71] Sanghyun Woo, Jongchan Park, Joon Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, volume 11211 LNCS, pages 3–19, 2018.