ALEXANDRA PEREIRA NUNES

UNDERWATER RECONSTRUCTION AND OBJECT RECOGNITION

SUPERVISOR: PROF. DOCTOR ANÍBAL CASTILHO DE COIMBRA MATOS



PH.D. THESIS

Doctoral Program in Electrical and Computer Engineering
2024, October

Alexandra Pereira Nunes
Underwater Reconstruction and Object Recognition
Ph.D. Thesis
2024, October.

E-MAIL:
apn@inesctec.pt

ABSTRACT

As ocean exploration is far from complete, it is crucial to develop methods that allow the Autonomous Underwater Vehicle (AUV) to properly perceive and respond to the observations. They must therefore overcome the inherent challenges of the underwater world, such as the difficulties of communication, as conventional methods do not work in water, and the pressure fluctuations at depth, which can deform systems and complicate the construction and maintenance of equipment. However, the biggest challenge in this area is the limited visibility, especially in deep and murky waters, which makes it difficult to interpret the surroundings. Nowadays, automatic perception of the underwater world is both challenging and crucial, and for this reason it is necessary to understand how this task can be made easier. It is known that the quality of visual data is the biggest challenge in analysing and understanding these scenarios, both for humans and robots. Acoustic data is an alternative as it can be used in all conditions, but it provides less information and is more difficult to interpret. It is therefore important to select the appropriate sensors in such a way that the advantages of each sensor can be fully utilised.

The main objectives of this work are therefore to develop methods for scene reconstruction using acoustic data that provide accurate representations of the robot's environment. In addition, learning techniques with training procedures will be developed that are capable of identifying, classifying, segmenting and extracting relevant information about the structural elements from the visual data.

Another important aspect of underwater perception research is the datasets that allow the methods developed to be tested objectively, taking into account different sensor configurations, environmental conditions, etc. However, in the underwater world there is currently a lack of suitable datasets for real-world conditions compared to other areas, as this collection is not easy and is an expensive task that requires suitable material. Therefore, it is often necessary to use simulators at the beginning of the experimental tests. They can replicate some conditions of the planned scenarios, namely some difficulties related to the underwater environment. In this case, the Stonefish simulator was used to simulate the seabed of a harbour and record acoustic images of Forward-Looking Sonar (FLS). An experimental setup was created consisting of a camera, a FLS and a positioning system that can be used to mimic the successive poses of the setup during a data acquisition mission. This setup can be mounted in a Remotely Operated Vehicle (ROV) so that real data acquisition is possible.

Therefore, an evaluation of the behaviour of the Mechanical Scanning Sonar (MSIS) during mapping tasks was performed in a test pool with some objects with different properties and building materials. Thus, 3D reconstructions were created using the sonar data with an external source for positioning. The results show that the shape of the objects can be correctly identified and the determined dimensions of the objects are well estimated with an error of less than 5 cm. Also in the field of reconstructions, a simple mosaic of acoustic images based on appearance was created, showing that it is possible to obtain a 2D map of the seabed with acoustic images of FLS. It allows a better visualisation of certain areas, focusing on the observation of objects that can be encountered, because the automatic detection of objects that could appear during the exploration is crucial for the robot to take preventive action. With objects such as a carabiner, a part of a tyre, a fine metal grid, a rope, an anchor, a plastic cup and a box, the results of the mosaic are well combined and the final result shows quality in terms of visual and easy perception of what is seen in the image. In terms of object classification, the developed Convolutional Neural Network (CNN) is evaluated with visual images under different conditions and compared with a state-of-the-art method. The results show that the developed method is suitable to classify 7 classes with good performance and achieves 75% under the most tested conditions. Next, a comparative study of suitable state-of-the-art semantic segmentation algorithms was conducted to determine whether they can be used in underwater scenarios. The SegNet model shows the best overall accuracy and achieves an accuracy of more than 80% when segmenting 8 classes and the mean value of Intersection-Over-Union (IoU) reaches 53%. However, some classes such as robots and plants degrade the quality of the performance, leading to results that can be dangerous in real-world contexts. In the last work in this field, an attempt was made to increase the performance achieved. Therefore, the application of some methods to improve image quality and data augmentation was compared. The results obtained with both strategies show that they are favourable, as they achieve a mean IoU of 56%. The result for each class shows that there are five classes with a IoU value of almost 60% and only one class with a IoU value of less than 30%, which is a more reliable result.

In this way, progress has been made in the reconstruction of underwater scenes and the classification of objects, thus improving the perceptual capabilities of the robotic systems.

Keywords: 3D Reconstruction; Mapping, Visual; Acoustic; Occupancy Grid; Object Classification; Machine Learning; Semantic Segmentation

RESUMO

Uma vez que a exploração dos oceanos se encontra longe de estar completa, é crucial desenvolver métodos que permitam aos veículos autónomos subaquáticos ter capacidades de perceção apropriadas, respondendo corretamente às observações. Devem assim superar os desafios inerentes ao ambiente subaquático, tais como as dificuldades de comunicação, uma vez que os métodos convencionais não funcionam na água. Além disso é necessário lidar com as flutuações de pressão em profundidade, que podem deformar os sistemas e complicar a sua construção e manutenção. Contudo, o maior desafio nesta área é a visibilidade reduzida, especialmente em cenários profundos e com turbidez, o que dificulta a interpretação do ambiente. Atualmente, a perceção automática do mundo subaquático é desafiante e, ao mesmo tempo, crucial. Desta forma, é necessário compreender como esta tarefa pode ser facilitada. Sabe-se que a qualidade dos dados visuais é o maior desafio na análise e compreensão destes cenários, tanto para os humanos como para os robôs. Os dados acústicos são uma alternativa, pois podem ser utilizados em todas as condições de visibilidade mas fornecem menos informações e são mais difíceis de interpretar. Assim, é importante selecionar os sensores apropriados por forma a aproveitar as mais valias de cada um.

Os principais objetivos deste trabalho são desenvolver métodos para reconstrução de cenas utilizando dados acústicos que forneçam representações precisas do ambiente em torno do veículo. Além disso, serão desenvolvidas técnicas de aprendizagem com procedimentos de treino que, a partir de dados visuais, são capazes de classificar, segmentar e extrair informação relevante sobre os objetos presentes.

Outro aspeto importante nesta área são os conjuntos de dados que permitem testar os métodos desenvolvidos de acordo com o objetivo, tendo em conta diferentes configurações de sensores, condições ambientais, etc. No entanto, neste ambiente há uma grande falha de conjuntos de dados em comparação com o meio exterior, pois esta recolha de dados não é fácil, sendo uma tarefa dispendiosa que requer material apropriado. Deste modo, muitas vezes é necessária a utilização de simuladores no início dos testes experimentais. Estas ferramentas podem replicar algumas condições dos cenários planeados, nomeadamente algumas dificuldades relacionadas com o ambiente subaquático. Neste caso, o simulador Stonefish foi utilizado para simular o fundo de um porto e adquirir imagens acústicas de um FLS. Foi também criado um sistema experimental composto por uma câmara, um FLS e um sistema de posicionamento que pode ser usado para

obter as sucessivas poses do sistema. Este sistema percetivo pode ser montado num ROV. Desta forma uma avaliação do comportamento do MSIS durante tarefas de mapeamento foi realizada num tanque experimental com alguns objetos que têm características e materiais de construção diferentes. Assim, foi criada uma reconstrução 3D utilizando os dados do sonar juntamente com uma fonte externa para posicionamento. Os resultados mostram que a forma dos objetos pode ser identificada e as dimensões dos objetos são bem estimadas com um erro menor que 5 cm. Ainda no domínio das reconstruções foi também criada uma técnica simples de mosaicking de imagens acústicas baseada na aparência, mostrando que é possível obter um mapa 2D do fundo do mar com imagens de um FLS. Desta forma, é possível uma melhor visualização de certas áreas, focando-se na observação dos objetos que são encontrados pois a deteção automática dos mesmos é cada vez mais importante durante missões autónomas. Alguns objetos como um mosquetão, parte de um pneu, uma grelha fina, uma corda, uma âncora, um copo de plástico e ainda uma caixa foram reconstruídos corretamente, com qualidade visual, permitindo assim uma interpretação mais fácil do que está na imagem. Em termos de classificação de objetos, uma CNN foi desenvolvida e avaliada em diferentes condições e comparada com um método do estado da arte. Os resultados mostram que o método desenvolvido é adequado para classificar 7 classes com bom desempenho, atingindo 75% de precisão na maioria das condições testadas. Em seguida, foi realizado um estudo comparativo de algoritmos de segmentação semântica existentes na literatura para determinar se estes métodos podem ser usados em cenários subaquáticos. O modelo SegNet apresenta o melhor desempenho, atingindo mais de 80% de precisão e um valor médio de IoU de 53% na segmentação de 8 classes. No entanto, algumas classes, como robôs e plantas, degradam a qualidade do desempenho, levando a resultados que podem ser perigosos em contextos reais. No último trabalho procurou-se aumentar o desempenho alcançado anteriormente. Assim, foram comparados a aplicação de alguns métodos para melhorar a qualidade da imagem e aumentar o número e diversidade de imagens. Os resultados obtidos com ambas as estratégias mostram que são boas escolhas atingindo uma média IoU de 56%. O resultado para cada classe demonstrou ainda que existem 5 classes com um valor IoU de aproximadamente 60% e apenas uma classe com um valor inferior a 30%, o que se torna mais confiável.

Desta forma, foram realizados progressos na reconstrução de cenas subaquáticas e na classificação de objetos, e assim as capacidades de perceção dos sistemas robóticos foram melhoradas.

Palavras chave: Reconstrução 3D; Mapeamento; Visão; Acústica; Grelhas de Ocupação; Classificação de Objetos; Aprendizagem Máquina; Segmentação Semântica

AGRADECIMENTOS

Para mim, esta tese de Doutoramento simboliza o culminar de um ciclo que, embora muito trabalhoso, me deu enorme gosto em realizar. É hora de agradecer àqueles que me apoiaram e ajudaram.

Gostaria de começar os meus agradecimentos pelo meu orientador Prof. Doutor Aníbal Matos, por toda a sua dedicação para que este resultado fosse possível. Por todo o empenho, incentivo e motivação no período referente ao Doutoramento, mas também nos últimos sete anos em que me acompanha. É incrível a quantidade e qualidade de conhecimento que me transmite e a discussão de ideias que possibilita, e ainda a liberdade de pensamento que me proporcionou autonomia, resiliência e perseverança. Sem dúvida que o seu apoio e orientação foram dois fatores preponderantes para o sucesso alcançado no encerramento deste capítulo. Ser-lhe-ei eternamente grata por todos os ensinamentos. Ao Prof. Doutor Nuno Cruz, pelas oportunidades que me possibilitou ao longo dos últimos três anos, enquanto Professor e Coordenador do CRAS, bem como todos os ensinamentos passados através do seu sentido crítico e que nos leva às mais proveitosas discussões do intelecto. Aos meus colegas do CRAS pelo apoio, em especial ao Jorge Barbosa por me ajudar durante a realização da componente de setup experimental, bem como pelo apoio e acompanhamento em campanhas de aquisição de dados. Gostava ainda de agradecer ao Prof. Doutor João Pavão pelo papel preponderante que assumiu desde o início do meu percurso enquanto aluna, e de seguida Engenheira, fazendo-me acreditar que bastava querer para alcançar os meus objetivos e partilhando uma ideia fundamental na Ciência: "*Um mau resultado, também é um resultado*".

Quero agradecer aos meus pais pelo apoio incondicional ao longo dos meus trinta e um anos de vida e por me possibilitarem as condições para que eu pudesse sempre aprender e querer mais! Ensinaram-me, acima de tudo, a correr atrás dos meus sonhos. Ao meu irmão Miguel, que acompanhou de perto, e com grande carinho, estes últimos anos, e me ensinou muitas vezes que questionar o que à partida já sabemos é das ferramentas mais valiosas na nossa vida. À minha tia Glória, que sempre olhou para este meu objetivo com muito orgulho, carinho, amor e paciência. Obrigada por serem os meus pilares e por me ensinarem todos os dias o significado da palavra família.

Ao meu namorado, Jacob Portela, que acompanhou toda esta jornada com muita paciência, compreensão e motivação, mesmo nos momentos em que eu desanimava. Obrigada por acreditares sempre em mim! Acredita que me ensinas todos os dias a querer ser mais,

não desistindo nunca dos meus sonhos, e tendo sempre completa consciência do meu valor.

À Rita, pela amizade que desenvolvemos ao longo destes 18 anos e por me apoiar na realização dos meus sonhos, e acima de tudo por também partilhar os dela comigo! Juntas fechamos vários ciclos da vida, e este não podia ser diferente. É um gosto imenso trabalhar contigo e ver-nos a crescer juntas, agora também a nível profissional. Juntas fazemos uma boa equipa! Relembro-te, mais uma vez, a frase de Ralph Emerson: "*Nenhum caminho é longo demais quando um amigo nos acompanha*". Agradeço também ao Bruno pela amizade ao longo dos anos, por toda a compreensão e apoio sempre que precisava.

À Ana, à Raquel, à Marta, ao Alexandre e ao Tiago, por continuarem a apoiar-me, após tantos anos, nesta loucura que sinto por aprender mais e mais, demonstrando interesse pelos meus objetivos e acreditando sempre que eu era capaz. À Catarina, ao João e à Luísa, por entrarem na minha vida e ajudarem a colori-la.

Obrigada a todos aqueles que cruzaram o seu caminho comigo e me ajudaram a evoluir dia após dia!

Alexandra Nunes

OFFICIAL ACKNOWLEDGEMENTS

Alexandra Pereira Nunes would like to thank the Portuguese Government through the FCT - Fundação para a Ciência e Tecnologia - and FSE - Fundo Social Europeu through the Norte 2020 - Programa Operacional Regional do Norte - for its doctoral scholarship SFRH/BD/146461/2019, without which this research would not have been possible.

In addition, Alexandra Nunes also like to thank the Centre for Robotics and Autonomous Systems (CRAS) of INESC TEC for being the host institution and also supporting this research.



Without continual growth and progress, such words as improvement, achievement and success have no meaning.

BENJAMIN FRANKLIN

CONTENTS

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Motivation . . . 1
- 1.3 Goals . . . 2
- 1.4 Problem Statement . . . 3
- 1.5 Contributions . . . 3
- 1.6 Thesis Overview . . . 5

i Fundamentals

2 Background 9
- 2.1 Underwater Perception . . . 9
- 2.2 Object Recognition and Mapping . . . 12

3 State of The Art 17
- 3.1 Underwater Scene Reconstruction . . . 17
- 3.2 Object Recognition . . . 24
- 3.3 Semantic Mapping . . . 32
- 3.4 Available Datasets . . . 34

ii Scene Reconstruction and Object Recognition

4 Tools and Robotic Platforms 41
- 4.1 Introduction . . . 41
- 4.2 Stonefish Simulator . . . 42
  - 4.2.1 Build Datasets . . . 45
- 4.3 Aquisition System . . . 48
- 4.4 Built Datasets . . . 58
  - 4.4.1 *CriticalObjectDataset* . . . 59
  - 4.4.2 WavesObjectsData . . . 61

5 Imaging Acoustic Map 63
- 5.1 Introduction . . . 63
- 5.2 Occupancy grid mapping from 2D sonar data . . . 64
  - 5.2.1 Methodology . . . 64
  - 5.2.2 Experimental Results . . . 70
  - 5.2.3 Discussion . . . 79
- 5.3 Mosaicking Approach with acoustic images . . . 80
  - 5.3.1 Methodology . . . 81
  - 5.3.2 Experimental Results . . . 86
  - 5.3.3 Discussion . . . 95
- 5.4 Summary . . . 96

6 Visual Object Recognition 99
- 6.1 Introduction . . . 99

6.2 Critical Object Classification 100
6.2.1 Methodology 101
6.2.2 Experimental Results 103
6.2.3 Discussion 107
6.3 Comparative Study of Semantic Segmentation Methods . 108
6.3.1 Methodology 109
6.3.2 Experimental Results 114
6.3.3 Discussion 123
6.4 Improving Semantic Segmentation Performance 124
6.4.1 Methodology 125
6.4.2 Experimental Results 133
6.4.3 Discussion 150
6.5 Summary 151

iii Final Considerations

7 Conclusion and Future Work 157

LIST OF FIGURES

1 Differences between the various tasks: Object identification, Object Classification and Object Segmentation. 13
2 Main phases of the process to obtain the semantic map approach. 15
3 UFO-120 dataset: real images and the corresponding ground truth about the existence of the objects. 36
4 SUIM dataset: real images and the corresponding ground truth masks with the labelled classes. 37
5 Some examples of acoustic images provided by *SoundMetrics*. 37
6 Example of the three maps used for the rendering process to create a terrain from the original image. 43
7 Example of the final result of the rendering process. 44
8 Example of the simultaneous acquisition of data by the camera and FLS in the simulator. 44
9 Example of the scenarios developed for the depiction of harbour facilities. 45
10 FLS model corresponds to the configurations during the acquisition of the dataset *Port_Mosaic*. 46
11 Examples of the acoustic FLS images obtained with the Stonefish simulator. 46
12 Trajectory of the vehicle during the *Port_Mosaic* dataset. 47
13 ROI that allows the selection of the part of the acoustic image that represents the returned echoes. 47
14 Main modules that are normally part of robotic vehicles. 48
15 Models of FLS from Tritech: Gemini 720im (a) and Gemini 720ik (b). 51
16 Model of FLS selected from Teledyne. 52
17 Model of FLS selected from Imagenex. 52
18 Some examples of chessboard images in which the inner corners are recognised regardless of the perspective. 54
19 Structured light system with a camera (a) and a line projector (b). 54

20 Scheme of the connections of the final perception system developed. 55
21 Customizable Remotely-operated Aquatic Bot. . 56
22 Mechanical Scanning Sonar used in the ROV. . . 57
23 Perceptual system installed on the CRAB vehicle. 57
24 Underwater GPS-G2 System. 58
25 An example of a buoy under both conditions: crystal clear a) and cloudy b). 60
26 Some examples of the critical objects in different acquisition conditions. 61
27 Some examples of the objects encountered during the WAVES campaign. 62
28 Setup of the camera mounted on the roof of the experimental pool for data acquisition. On the right you can see the image provided by one of the cameras. 65
29 Example of a marker used (a) and the reference coordinates (b) of the determined position. . . . 66
30 Result of the position achieved by marker tracking (x, y, yaw). 66
31 Setup configuration for data acquisition with an ArUco marker and Tritech MSIS. 67
32 Overview of the developed point accumulation method. 68
33 Probability assigned to each sample according to the scan angle when the sonar is configured with an angle of -15° to 15°. 70
34 The 5 types of material used to evaluate the sensor: wood (a), PVC (b), metal (c), concrete (d) and plastic (e). 71
35 A metal box a) and a plastic bottle b) were recognised with a gain of 20% in the software of MSIS. A plastic bottle with 70% gain is also shown c). 71
36 Objects used to test the developed method: a box (a), a boat hull (b) and a trunk (c). 73
37 The original yaw angle a) was filtered to ensure consistency along the trajectory b). 73
38 Final map of the surface of the trunk and verification of the dimensions in relation to the height. 75
39 The final map shows the shape of the hull (top) and the verification of the dimensions in x and z coordinates (bottom). 76
40 Final reconstruction of scenario E (top) and verification of the dimension in the y coordinate - width of the boat (bottom). 77

41 Discretisation process applied to scenario E with a step size of 5cm. 78
42 Illustrative example of a mosaic after the processing of 3 images. 80
43 Pipeline of the implemented mosaic method for the application to FLS images. 81
44 The different regions of the *Port_Mosaic* Dataset: sand (a), metal grid (b) and stones (c), presenting different textures. 83
45 The differences between the matches obtained with ORB extractor (a) and BRISK extractor (b). 84
46 Example of some finished mosaics with areas with black pixels that lead to an unsightly result. 86
47 Illustrative representation showing the introduction of the cropping method when combining three images. 86
48 Example of a mosaic with two images in the stone area with a forward movement. 87
49 Combination of 15 consecutive images (a) and their movement according to the ground truth (b). 88
50 Example of a mosaic of 2 images with 10 good matches (a) and 29 good matches (b) caused by different configurations. 88
51 Differences between the merging of 2 images with: 500 (a), 1500 (b), 3000 (c), 4000 (d) and 5000 (e) extracted features. 89
52 Final result obtained by combine 6 images of the different region: sand (a) and grid (b). 90
53 Individual acoustic images of a carabiner (a) and the result of the combination of successive images of this object (b). 90
54 Individual images (acoustic and corresponding visual) of the tyre part (a) and two examples of results obtained by combining individual acoustic images of this object (b). 91
55 Results obtained where a rope (a) and a fine grid (b) can be observed. 92
56 Single visual image of the anchor (a) and the result of combining consecutive acoustic images (b). 92
57 Differences on the obtained mosaic extracting more features (6000), but using different requirements on the test ratio: 0.8 (a) or 0.9 (b) for maximum distance. 93

58 Result with a plastic cup (a) and a box (b) with 6 consecutive images of each object. 94
59 Single acoustic image of the vase (a), the result of combining single acoustic images with FOV of 40° and with FOV of 50°. 94
60 The 7 critical objects used in the *CriticalObject-Dataset*. 101
61 Process overview of the proposed CNN method. 102
62 The performance achieved in both phases: training and validation, as well as the loss value during 50 epochs. 104
63 Achieved performance in terms of accuracy for each class over 50 epochs. 104
64 SUIM dataset: real images and the corresponding ground truth masks. 113
65 Distribution of the individual classes for the images of the original training set. 114
66 Difficult images to test the algorithms: noise, light attenuation and many classes per image. . 114
67 Segmented image after 30 evaluations by SegNet (a), PSPNET (b) and FCNN_16s. 116
68 Evolution of the results in the learning process before: 200 training iterations (a), 1500 training iterations (b), 15000 training iterations (c), 45000 training iterations (d), 75000 training iterations (e) and 100000 training iterations (f). 116
69 Evolution of the results in the learning process of the original image (a) before 15000 training iterations (b), before before 45000 training iterations (c) and before 100000 training iterations (d). 117
70 Comparative performance results at different learning rate. 120
71 Comparative performance results using different weight decay values. 120
72 Performance evaluation over 200000 training iterations, with evaluations after every 500 iterations. 121
73 Illustrative images of the dataset without robot class. 123
74 Diagram of the general framework of the proposal for improve the semantic segmentation results. 125

75 The segmented images obtained with the original dataset for a total of 100 evaluations after 500 training images each, based the best model. 126
76 Problems of the SUIM dataset: more elements of a class in the original image than in the mask (a), some labeled errors when referring to the same class (b) and poor quality of the images (c). 127
77 Comparison of the improvement methods using 6 original images (a) with the CLAHE method (b), white balance transformation (c) and dive correction (d). 129
78 Five examples of possible changes to increase the size of the dataset using a single image. . . . 131
79 Result from two original images (a): in some cases, good variations result, but at the same time images arise that do not fit into the intended context (b). 131
80 Some examples of original images (a) and corresponding transformations (b). The first and second rows represent transformations by flipping and cropping and the last row represents the deterioration in quality. 133
81 Quality assessment with BRISQUE and PIQUE for 20 original images selected to evaluate the quality of images processed with enhancement techniques. 134
82 Two examples of bad visual images (a) and good visual images (b) of the original set, according with the methods BRISQUE and PIQUE. 135
83 Qualitative comparison results of 4 different examples (a and b): the CLAHE method (c), WB method (d) and dive correction (e) using the best model in a total of 100 validations, each after 500 training images. 137
84 Comparison of the qualitative results of 2 different examples (a and b), between the result obtained with the original set (c) and the CLAHE method (d) using the best model for a total of 100 validations. 138
85 Eight transformations used in the creation of the augmented datasets for training. 139
86 Distribution of the individual classes for the images of the training set with Data Augmentation. 139
87 Mean IoU measurement along the 50 evaluations with the original dataset and the DA dataset. 141

88 IoU measurement along the 50 evaluations with the original dataset and the DA dataset, for each class. 141
89 Qualitative comparison of the results of 10 different examples of semantic segmentation (a and d) using the original dataset (b and e) and DA set (c and f) for the training process, where the best model was selected in a total of 50 validations. 142
90 Distribution of the individual classes for the images of the training set when applying the data augmentation in different number according to the class. 142
91 Qualitative comparison results of 4 different examples of semantic segmentation (a and b), obtained from the original training set (c) and the DA_v1 dataset (d), using the best model in a total of 50 validations for each case. 144
92 Distribution of the individual classes for the images of the training set when applying the data augmentation only in the robot class. 145
93 IoU measurement for the robot class (class 2) along the 50 evaluations with the original dataset and the Data Augmentation dataset. 146
94 Qualitative results of different examples of segmented image with robots, obtained from the Data Augmentation set for the robot class, using the best model in a total of 50 validations. 146
95 Demonstration of the size and location on the original set of the plants in some of the images. . 146
96 Distribution of the individual classes for the images of the training set when applying the Data Augmentation only in the plant class. . . . 147
97 IoU measurement for the plant class (class 4) along the 50 evaluations with the original dataset and the data augmentation dataset. 147
98 Results of 5 different examples of semantic segmentation (a and b), obtained from the use of both strategies at the same time: data augmentation of the less representative classes and CLAHE method. 149

LIST OF TABLES

1 Comparison of some perceptual sensors for underwater applications. 11
2 Advantages and disadvantages of the various traditional methods of object recognition. 24
3 Comparison between the 5 selected models of FLS to analyse their main characteristics. 52
4 Comparative analysis of the number of points without and with the filter techniques. 70
5 Results in relation to the returned echoes with variation of the gain parameter. 72
6 The scenarios used to test the developed implementation. 73
7 Dimensions obtained of the metal box in scenario A. 74
8 Determined dimensions of the trunk in scenario C. 74
9 The number of points after the discretisation process for scenario E. 78
10 Average number of keypoints and matches in 10 consecutive images in three different regions: Sand, Grid and Stones. 83
11 Accuracy metrics and time required when using the proposed method and ResNet_v1 for the *datasetA*. 105
12 Accuracy metrics and time required using the proposed method and ResNet variations for the *datasetB*. 105
13 Accuracy metrics and time required when using the proposed method and ResNet_v1 for the *datasetC*. 106
14 Accuracy metrics and time required when using the proposed method and ResNet_v1 for the *datasetD*. 106
15 Accuracy [%], IoU [%] and time results after 100000 training iterations. 115
16 IoU [%] per class for 66 evaluations after 100000 training iterations. 117
17 Accuracy [%], IoU [%] and time results when varying the number of workers. 118

18 Accuracy [%], IoU [%] and time results varying the size of the training images. 119
19 IoU [%] per class after 100000 training iterations, if the evaluation is carried out after every 500 training images (A) and if the evaluation is carried out after every 1500 training images (B). 121
20 Performance results [%] after 100000 training iterations without robot class. 123
21 Performance measurements [%] for the best result obtained with the original dataset in a total of 100 evaluations. 126
22 Comparative results of the mean of the quality score of the Dive, WB and CLAHE corrections for a total of 20 images, using PIQUE, BRISQUE, NIQE and the *sr_metric*. For each metric, the results that exceed the results obtained with the original set are highlighted in bold. 135
23 Comparative results of best measured model performance [%] in terms of Accuracy (Overall and Mean) and IoU (Mean) achieved with: CLAHE, WB and dive corrections in a total of 100 validations, each after 500 training images. Whenever the results achieved exceed the results of the original set, they are emphasised in bold. 136
24 Comparative results of best measured model performance [%] in terms of IoU for each class achieved with: CLAHE, WB and dive corrections in a total of 100 validations, each after 500 training images. Whenever the results achieved exceed the results of the original set, they are emphasised in bold. 136
25 Performance measures [%] for the best result obtained with the original dataset and the dataset with DA in a total of 50 validations. In bold type you can see that the results achieved with the data augmentation are better than the original results. 140
26 IoU [%] for each class in the best result obtained with the original dataset and the dataset with DA in a total of 50 validations. In bold type you can see for each class whether the results achieved with the data augmentation are better than the original results. 140

27 Performance measures [%] for the best result obtained with the original dataset and the new (DA_v1) dataset in a total of 50 validations. In bold type you can see that the results achieved with the data augmentation are better than the original results. 143
28 IoU [%] for each class in the best result obtained with the original dataset and the new with data augmentation (DA_v1) in a total of 50 validations. In bold type you can see for each class whether the results achieved with the data augmentation are better than the original results. . . 143
29 Performance metrics [%] comparison in terms of Accuracy (Overall and Mean) and Mean IoU for the best result with the more balanced dataset with data augmentation (DA_v1) and this dataset with CLAHE method (DA_C) applied for a total of 50 validations. In bold type you can see that the results achieved with the data augmentation are better than the original results. 148
30 IoU [%] for each class in the best result with the more balanced dataset with data augmentation (DA_v1) and this dataset with CLAHE method (DA_vC) applied for a total of 50 validations. In bold type you can see for each class whether the results achieved with the data augmentation are better than the original results. 148

ACRONYMS

ANFIS Adaptive Neuro-Fuzzy Interference System.
ASV Autonomous Surface Vehicles.
AUV Autonomous Underwater Vehicle.

BRIEF Binary Robust Independent Elementary Features.
BRISK Binary Robust Invariant Scalable Keypoints.
BRISQUE Blind/No-Reference Image QUality Evaluator.

CLAHE Contrast Limited Adaptive Equalisation.
CNN Convolutional Neural Network.
ConvNet Convolutional Network.
CRAB Customizable Remotely-operated Aquatic Bot.

DA Data Augmentation.
DVL Doppler Velocity Log.

FAST Features from Accelerated Segment Test.
FCNN Fully Convolutional Networks.
FLANN Fast Library for Approximate Nearest Neighbors.
FLS Forward-Looking Sonar.
FOV Field-of-View.
FSS Forward-Scanning Sonar.

GigE Gigabit Ethernet.
GPS Global Positioning System.
GPU Graphics Processing Unit.

IMU Inertial Measurement Unit.
INS Inertial Navigation System.
IoU Intersection-Over-Union.
IQA Image Quality Assessment.
IS Imaging Sonar.

KNN K-Nearest Neighbour.

MBS MultiBeam Sonar.
ML Maximum Likelihood.
MSIS Mechanical Scanning Sonar.
MUED Marine Underwater Environment Database.

NIQE Natural Image Quality Evaluator.

OpenCV Open Computer Vision Library.

ORB Oriented FAST and Rotated BRIEF.

PIQUE Perceptual-based Image QUality Evaluator.
PS Profiling Sonar.
PSPNET Pyramid Scene Parsing Network.
PT Processing Time.

RANSAC Random Sample Consensus.
ResNet Residual Neural Network.
ROI Region of Interest.
ROS Robot Operating System.
ROV Remotely Operated Vehicle.

SIFT Scale-Invariant Feature Transform.
SLAM Simultaneous Localization and Mapping.
SSS Side-Scan Sonar.
SUIM Segmentation of Underwater IMagery.
SV Stereo Vision.
SVM Support Vector Machines.

TA Training Accuracy.
TeA Test Accuracy.

VA Validation Accuracy.
VIO Visual-Inertial Odometry.
VO Visual Odometry.

WAVES Advanced Vehicles for Exploration of the Seas Workshop.
WB White Balance.

YOLO You Only Look Once.

1 INTRODUCTION

1.1 CONTEXT

The underwater scenario is unexplored, which means that it is an open world, as the structures and marine life that exist there are still unknown. The exploration of the oceans is therefore far from complete and the AUV can help in various fields such as marine biology, geology and archaeology, namely ensuring that divers do not get into dangerous situations. This environment is not only unknown but also unstructured and therefore these vehicles must be able to better perceive the environment to be explored, i.e. obtain reconstructions of the surrounding scenes and identify some objects that might appear during the mission. In addition, it is still a big challenge to collect high quality data from visual and acoustic sensors to understand the scenes well, which makes this task even more difficult. The challenges in the underwater world are great because the conditions are unique and very different from those outside, especially in terms of the quality of the data collected. The perceptual information collected in this environment is poor, mainly due to the limited visibility. Therefore, some efforts must be made to select the most suitable acoustic and visual sensors and to develop methods that improve the quality of the data and thus the result of their processing.

1.2 MOTIVATION

As ocean exploration is far from complete, it is crucial to develop methods that enable vehicles to have autonomous perception capabilities and respond correctly to observations. They must therefore overcome the inherent challenges of the underwater environment to gain an accurate and high-quality understanding of the vehicle's surroundings. To fulfil these tasks underwater, it is often important to use the information provided by the acoustic sensors or the visual systems, but there are many differences between them. While the visual sensors provide a lot of information but are of limited use in areas with poor visibility, the acoustic sensors allow use without these disturbances but are more difficult to interpret and their behaviour changes with the water conditions. It is therefore important to develop methods that can accurately reconstruct scenes and provide both metric measurements and detailed information. In addition, the diversity of objects in the

sea is another motivation for this study, as there are different types of objects in this environment (natural, man-made, critical for the operation of vehicles, floating, etc.). The methods developed must therefore be able to recognise each intended class and make decisions so that the vehicle reacts in the best way depending on the type of object. This task is crucial for other more complex tasks such as object manipulation or obstacle avoidance. The methods that are able to reconstruct the scene require some processing time, but the main goal is that some important information about the environment is extracted in a simple way. In the field of object recognition, machine learning methods have proven their potential. To train a machine learning method correctly, a large number of features based on different information is required. There is also a lack of available datasets and the existing ones are incomplete as data collection is more costly in this environment. So, there is a need for underwater data as a large number of previous observations are required for the development and evaluation of perception methods. Therefore, it is necessary to explore both topics in order to advance technology and improve the perceptual capabilities of AUV.

1.3 GOALS

The main focus of this doctoral thesis is to improve the perception of vehicles in the still unknown underwater environment. The idea is that they can operate more and more without human control, even in unknown, unstructured and sometimes inaccessible and dangerous regions. The main goals of this work are therefore:

1. To improve the perceptual capabilities of the robotic vehicle so that it can better perceive the unknown scenes during navigation. The effects of the individual sensors on the control and navigation of the robot must be evaluated in order to select a suitable experimental setup.

2. To allow the vehicle to reconstruct the scene that surrounds it in an unknown and unstructured environment. In addition, a more comprehensive survey and inspection mission can be performed to provide the user with further information that could be of crucial importance, such as the condition of the structures or the size of the objects that could appear in the reconstruction.

3. To contribute to the vehicle's ability to correctly identify and classify the objects in the environment during navigation in order to maximise the perceptual quality of the other tasks that require this information. It also increases the autonomy of vehicles, as they can perform certain actions depending on the object they encounter.

1.4 PROBLEM STATEMENT

Therefore, the methods of scene reconstruction using acoustic data are explored in an accurate way. Furthermore, learning techniques with training procedures and the validation of artificial intelligence agents capable of identifying, classifying and extracting relevant information about the structural elements from the visual data are investigated. The main objectives of this work are therefore:

- To evaluate the perceptual needs and limitations of the underwater world. Also analyse a perceptual solution for acoustic and visual information suitable for the development of the proposed work.
- To develop methods to reconstruct the round scenario based on acoustic data during an inspection.
- To develop methods that make it possible to categorise the objects in the scene using artificial agents - supervised learning.
- To collect real data in an underwater environment with various sensors to check the quality of the developed methods and prove their feasibility.

1.5 CONTRIBUTIONS

As the main contribution of thesis are:

- Selection and construction of a system for real data acquisition, consisting of a camera, a FLS and a ground-truth localisation system.
- Development of a method to reconstruct the scenes under the vehicle using a MSIS and an external position source.
- Development of a method that allows a mosaicking to be created from FLS images using visual assumptions.
- Development of a method for classifying critical objects in underwater optical images that could be dangerous to humans in the underwater world.
- Comparative study and needs analysis of semantic segmentation methods that can work with underwater images.
- Implementation of strategies that increase the performance of the semantic segmentation method to use these techniques with accuracy in a real context.

Publications

- **Improving Semantic Segmentation Performance in Underwater Images.**

  Alexandra Nunes, Ana Rita Gaspar and Aníbal Matos.

  Journal of Marine Science and Engineering, 2023

  (10.3390/jmse11122268)

  Nowadays, semantic segmentation is increasingly used in underwater applications, but it is necessary to improve the performance in these tasks. Therefore, the application of some methods to improve image quality and data augmentation were compared. The results obtained with both strategies show that they are favourable, achieving a mean IoU of 56%. The result for each class shows that there are five classes with an IoU value of almost 60% and only one class with an IoU value of less than 30%, which is a more reliable result.

- **Comparative Study of Semantic Segmentation Methods in Harbour Infrastructures.**

  Alexandra Nunes, Ana Rita Gaspar and Aníbal Matos.

  OCEANS 2023 Limerick, 5-8 June

  (10.1109/OCEANSLimerick52467.2023.10244570)

  A comparative study of suitable state-of-the-art semantic segmentation algorithms is conducted to determine whether they can be used in underwater scenarios. Therefore, several parameters and different conditions have been considered to perform a complete evaluation and some limitations and improvements are described. The SegNet model shows the best overall accuracy, reaching more than 80%, but some classes such as robots and plants degrade the quality of the performance.

- **Occupancy Grid Mapping from 2D SONAR Data for Underwater Scenes.**

  Alexandra Nunes, Ana Rita Gaspar and Aníbal Matos.

  OCEANS 2021 - San Diego - Porto, 20-23 September

  (10.23919/OCEANS44145.2021.9705764)

  There is a preliminary study to evaluate the performance of the imaging multibeam sonar in mapping tasks. The evaluation is performed by some experiments on a test pool with some objects with different properties and construction materials. So, a 3D reconstruction is created using the sonar data with an external source for positioning. The results show that the shape

of the objects can be correctly identified, and the determined dimensions of the objects are well estimated.

- **Critical object recognition in underwater environment.**

  Alexandra Nunes, Ana Rita Gaspar and Aníbal Matos.

  OCEANS 2019 - Marseille, 17-20 June

  (10.1109/OCEANSE.2019.8867360)

  A dataset was developed with some man-made critical objects. The developed Convolutional Neural Network is presented and its evaluation under different conditions is carried out. A comparative analysis and a discussion between the results of the proposed algorithm and the results of a state-of-the-art architecture are also presented. The results show that the developed method is suitable to classify 7 classes with good performance.

1.6 THESIS OVERVIEW

This dissertation is divided into 3 parts and 7 chapters. The "**Fundamentals**" part consists of chapters 2 and 3.

- Chapter 2 provides a better understanding of the challenges of underwater perception and explains how each sensor can help with data collection. It then introduces some concepts of identification, classification and semantic mapping to better understand the issues involved.
- Chapter 3 describes the most important work in the literature on environment reconstruction and object classification and some analyses of the results obtained. It summarises some advances in semantic mapping and describes the current state of the art in terms of available datasets.

The second part "**Scene Reconstruction and Object Recognition**" consists of the chapters 4 to 6 and describes the practical part of this thesis.

- Chapter 4 presents the advantages of using simulations to test the developed methods in a preliminary way, some crucial questions and the developed datasets in the selected tool. It also compares the sensors best suited for good underwater perceptual data collection and the experimental setup used to collect datasets in the real environment.

- Chapter 5 presents a simple method that combines the information obtained from an imaging sonar with an external information source and demonstrates its behaviour when mapping some objects above the sensor. Furthermore, a method is presented that combines acoustic images with the same assumptions as visual mosaicing based on appearance, as well as their evaluation in simulated data.

- Chapter 6 presents a preliminary study on critical object recognition, comparing an in-house implementation with one from the literature, using a previously developed dataset. A comparative study of different literature approaches to image segmentation is then applied to a known underwater dataset and then some strategies to improve the previously obtained results are explored.

The last part, "**Final Considerations**", consists of chapter 7, where the main conclusions of the whole thesis are discussed and some ideas for future work are presented.

Part I

FUNDAMENTALS

2 BACKGROUND

In most cases, exploring the underwater world with robotic vehicles requires methods for automatically observing and analysing the information provided by the perception systems. However, this information must be presented in the best possible quality in order to be processed correctly. Therefore, this chapter presents some difficulties related to underwater exploration with perceptual sensors and some concepts related to the methods analysed in this thesis.

The chapter is organised as follows: section 2.1 provides a better understanding of the difficulties associated with underwater perception and explains how each sensor can help collect perceptual data and when it should be used. Section 2.2 then introduces some concepts of object identification, classification, segmentation and semantic maps.

2.1 UNDERWATER PERCEPTION

The challenges in the underwater world are great because the conditions are unique and very different from those outside. The biggest challenge in this area is the limited visibility, especially in deep and murky scenarios, which can be dangerous for autonomous navigation. The difficulties in communication, as conventional methods do not work in the water, and the pressure fluctuations at depth, which can deform the systems and make the construction and maintenance of the devices more difficult, are also challenges that need to be considered. In addition, there is a lot of marine life in the water and it is important that exploration is carried out without disturbing them. Advances in technology and ongoing research are therefore helping to improve the results obtained in this study and make them more accessible.

Nowadays, the automatic perception of the underwater world is both challenging and crucial, and for this reason it is necessary to understand how this task can be facilitated. It is known that the quality of the data is the biggest challenge in analysing and understanding these scenarios, both for the robots and for the humans. The acoustic data is an alternative as it can be used in all conditions (poor visibility, murky environment, etc.). Otherwise, visual sensors are always better suited for use at close range [1]. The optical sensors can provide high-resolution images that are easy to interpret, but their range is limited to a few metres. Therefore, the use of cameras in this environment is often not possible due to unfavourable environmental conditions, namely

light absorption and scattering. In addition, poor control of data capture, insufficient lighting and limited accessibility make the use of cameras considerably more difficult. Nowadays, some processing techniques are already used to improve the quality and contrast of the image, which is called image enhancement. In many cases, this enhancement is the only way to utilise the underwater images. It is usually necessary to correct the vignetting effect (the edges of the image are darker than the centre) or colour correction caused by light attenuation [2]. It is also possible to apply some techniques to reduce noise or even highlight some features of the images that are less visible. In many cases, improving image quality requires a more accurate knowledge of the model to estimate water absorption and scattering with the aim of removing haze [3]. But often the solution is not sufficient to ensure the quality for extracting perceptual information or it requires high computational costs. In 2017, a comprehensive review of optical image processing was presented with more than 40 papers on image enhancement that could help in many types of applications in the future [4].

These limitations are overcome by resorting to acoustic sensors, which are interesting because they are able to provide information in underwater environments, namely in poor or no visibility [5]. They cover a large area and allow data acquisition in real time with high frame rates, so that large areas can be mapped quickly. In this way, it is possible to capture the scenes without physical contact with the objects. For this reason, they are often used in tasks such as obstacle avoidance or object identification by robots or even in rescue operations and underwater archaeology by humans. These sensors obtain their measurements as a function of the speed of sound along the trajectory of the vehicle or sensor. However, the data provided by these sensors has a low resolution and the information they provide is limited as they do not provide colour information, which is relevant for many perceptual tasks [6]. Often these sensors are susceptible to acoustic interference and their behaviour depends on water conditions such as salinity or temperature of the water [7]. In addition, the processing of this type of data is more complex and their interpretation is not intuitive due to some effects of acoustic reflection and backscattering that also lead to geometric inaccuracies. In general, the cost of these sensors is higher than for other visual sensors. In the field of sonar technology, there are different ways to collect data: with one sensor or by combining two or more sources of information. In [5], an example of the combined use of high-resolution images from Side-Scan Sonar (SSS), which is better suited for operations over long distances, with data from multibeam FLS was presented. The latter allows investigation closer to the seabed and is therefore suitable for more detailed inspection. In many cases, the SSS is used to detect possible object positions and then a new and

more detailed search is performed in the FLS data to analyse if the objects of interest are really at these positions and to possibly perform some intervention actions. Furthermore, the sonar sensors can be active or passive, but regardless of the type, the data from the sonar depends on the depth of the reflecting surface [8]. Once the data has been collected, a visualisation of the scene can be created. However, this requires knowledge of the physical processes involved. For many tasks involving the perception and use of images, for example, the methods can be applied to any image to obtain the results, but when it comes to 3D reconstruction of scenes, it is a bit different. Nowadays, there are several alternatives when it comes to choosing sensors for this task. However, many of these sensors have only been tested under aerial conditions, so new approaches are needed for their application in the underwater environment. Therefore, the different solutions that enable 3D data have been categorised according to the technology used in [9]. Three main groups are thus distinguished: triangulation, time-of-flight measurement and modulation. The different sensors can belong to more than one class, depending on which method is used to obtain the three-dimensional data. Triangulation methods solve the problem by measuring the distance of two or more devices with a common feature or target, but they have the limitation that there must be an area of overlap between the devices. Time-of-flight methods are based on the time it takes for the signal to travel long distances, but can be affected by changes in the salinity or temperature of the water. Finally, frequency domain modulation uses the differences in amplitude and phase of a modulated signal to accomplish this task, but only the low-frequency signals are transmitted efficiently (the high-frequency signals are lost).

Table 1 contains a comparison of some perceptual sensors: Multi-Beam Sonar (MBS), Imaging Sonar (IS), SSS and Stereo Vision (SV), with its strengths and weaknesses.

Table 1: Comparison of some perceptual sensors for underwater applications.

Sensor	Range	Strengths	Weaknesses
MBS	11 km	- Independent of water turbidity - Long coverage	- High minimum distance
IS	150 m	- Independent of water turbidity - Extraction features information	- Low frame rate - Unknown dimensions
SSS	150 m	- Independent of water turbidity - High frame rate	- Requires constant speed - Unknown dimensions
SV	3m	- Simple - High accuracy - Provides colour information	- Low frame rate - Sparse data coverage - Affected by water turbidity

From this it can be concluded that the choice of the final system for the perception of the underwater world is not trivial. Otherwise, the weight, the volume of the payload, the measurement time, the budget and the availability of energy are aspects that should also be taken into consideration.

2.2 OBJECT RECOGNITION AND MAPPING

The classification of objects is an important task in the field of robotics, where in many cases images are the most important resource. As already mentioned, applying processing techniques to underwater images cannot provide good results because the feature information is very low compared to aerial images, the images are blurred and the colours are affected by the attenuation of light. In this environment there are both artificial and natural objects. The latter are more difficult to recognise as they resemble the seabed (i.e. can be masked) and the surroundings are unknown. In many cases, the objects are invisible on the surface of the water. Therefore, the advance information about the objects is helpful in the underwater search by the autonomous vehicle and its application prevents some dangers to humans. Otherwise, there could be objects in this environment that are under tension or buried, which also complicates the classification process. In order for the vehicles to fulfil this task, they must therefore be equipped with intelligence to react to the collected data.

In the context of image classification, there are three different concepts that are often confused with each other: identification, classification and segmentation. Identification refers to the process of recognising the existence of an object in an image (distinguishing objects from the background). Classification is the process of not only identifying an object, but also assigning it to a class or category. An entire object is assigned a label based on its characteristics or properties. Finally, segmentation is the process of dividing the image into segments that define the boundaries of different regions or objects within the scene. This is a dense classification as each pixel is assigned a class. To better understand the main differences between the topics, see Figure 1. However, the success of these three different tasks is influenced by many factors that can affect their good performance [10]. Some of them are:

- rotation - the object can remain in different rotated shapes on the images;
- positioning - the object can take different positions;
- occlusion - when the object is partially obscured or obscured by others;

- mirroring - the object can appear mirrored on the images;
- scale - the object can be displayed in different sizes depending on the distance between the camera and the object.

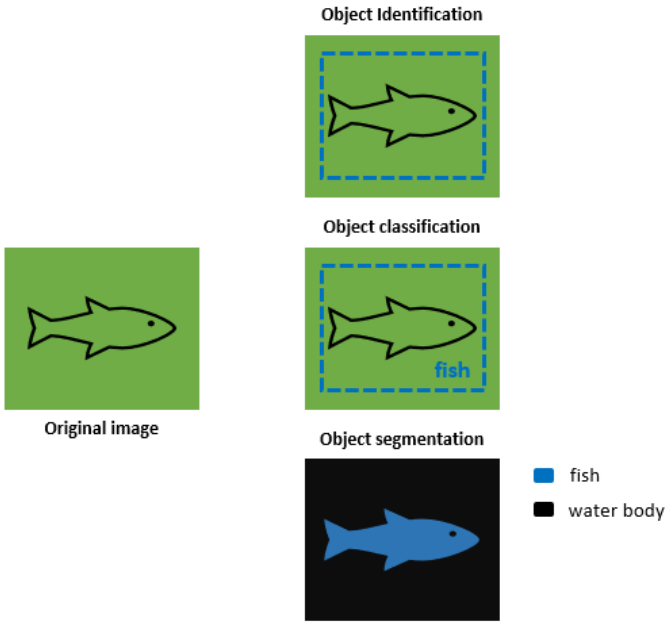


Figure 1: Differences between the various tasks: Object identification, Object Classification and Object Segmentation.

It is also possible to differentiate the methods of object identification by using a single image or by using temporal information computing a sequence of single images to reduce the errors and thus increase the accuracy rate. In addition to the difficulties of underwater image processing techniques, there are some problems that arise when applying classification methods in this environment, namely the fact that most of the current methods for object recognition are not able to deal dynamically with different seabeds, as this depends on where the survey takes place. Furthermore, the algorithms often do not take into account the physical and geometric knowledge of the objects.

If we want to move from the identification of objects to the classification of the different elements, conventional techniques are often limited when using raw data from different sensors, as already mentioned. Therefore, some learning techniques have been used in this area to solve this task in a simpler way. Better visual perception requires building representations with the quality of the real world. A frequently asked question in these applications is therefore: *"How to learn the most important features automatically?"*. These features must be able to recognise the objects regardless of their scale or position and classify them into categories. In image processing, there is an important hierarchy of elements in terms of complexity: pixels, motifs, parts, objects and scene [11].

In most cases, it is necessary to collect data to perform the training phase, since most of the methods used are unsupervised and require a large amount of data. In addition, some data is needed to test the developed method under conditions that are as close to reality as possible. It is important to point out that this is a difficult task, especially when the environment is underwater. Again, the success of the object classification depends heavily on the quality of the data to be used. Moreover, it is known that the learning techniques require a large amount of data to perform the training and consequently learn the most important features of the objects. In this way, it is possible to obtain multiple levels of representation that reinforce the aspects of the inputs that are crucial to distinguish the classes and suppress irrelevant variations, such as lighting, pose, background or environment. In recent years, new important studies have been carried out on the application of learning subsystems with feature extractors that transform the raw data into a suitable internal representation, since the success of the classification is often related to the success of the feature extractor [12]. Over time, the quantity of images and the number of categorised classes increase and the quality of the data improves. It is important to note that datasets such as ImageNet (explained in more detail in section 3.4) help with training without overfitting, i.e. the model fits the training data well but is not able to predict new results. In this way, these datasets were used to test the developed methods that allow the classification of more than two classes, for example with a Deep Convolutional Neural Network (CNN). The biggest advantage of the dataset is the large amount of labelled data. However, as object recognition is a very complex task, the number of features of each object cannot always be fully determined, even with a large data set. The method must therefore have a lot of prior knowledge to compensate for this fact [13]. If the computational requirements are not relevant, it is possible to use a CNN with a higher number of layers. Otherwise, using the new Graphics Processing Unit (GPU) makes this task less complex and the training phase is performed in a simple way. In addition, there are some techniques to reduce overfitting, such as:

- **Data augmentation** – this means that the data set is artificially enlarged (filtering of images, cropping, resizing, ...);
- **Dropout** – this means that not all units are used during the training phase (in general, the neurons used are selected at random)

As work progresses, more and more complex tasks are performed automatically and by robots, such as image segmentation, whose main objective, as already explained, is to assign each pixel of an image to a corresponding class of the displayed image. This is a dense

classification and it is often difficult to achieve good prediction results. There are several problems in this context, but the most important one is to achieve a balance between the accuracy and efficiency of semantic segmentation. Furthermore, it is a task that requires a complete data set that is appropriately labelled. This is difficult to achieve in many cases because if we want to create a dataset for this purpose, we need a labelled ground truth task, which in many cases is created manually. Another factor is the small number of datasets, which usually do not include all sizes, perspectives and scales of objects. In addition, data augmentation strategies are crucial to adapt the dataset to the different training data [14]. Therefore, it is possible that we enable the model to learn these variations during the training process and improve its performance when faced with such challenges. This technique is used in many contexts, but is very important in the underwater domain as the datasets are often small. Another strategy is the use of adaptive models. These models can dynamically adapt their behaviour to the characteristics of the input data, etc. It is crucial that this task has a higher accuracy as it is more and more integrated as part of other, more complex tasks such as manipulation or obstacle avoidance and the success of the big task has a lot to do with the success of semantic segmentation.

Nowadays, it is becoming increasingly important for robots to understand their environment and also be able to navigate autonomously, create semantic maps and, in some cases, avoid obstacles. In this way, robots are able to understand the world better and better. Such semantic information can enable a robot to better distinguish low-level locations or objects with similar geometric and visual characteristics and perform higher-level tasks that use semantic information about objects to manipulate and environments to navigate. So the decision levels can be higher because the robots can answer two different questions: *"Where am I?"* and *"What is this place?"* [15]. There are different ways to combine both pieces of information, but in Figure 2 you can see a representation that summarises the main idea.

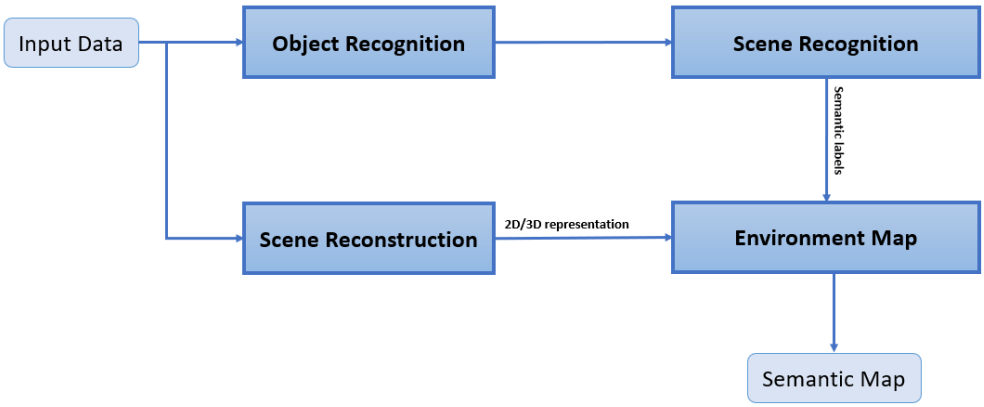


Figure 2: Main phases of the process to obtain the semantic map approach.

The input data is therefore sent simultaneously to the object recognition system and to the method for reconstructing the scene. Once the scene has been fully recognised, this information can be sent to the system, which combines it with the reconstruction to create the semantic map. The semantic information can be combined with many different techniques when reconstructing the scene so that the robot can create a map of the environment and localise itself in this map at the same time. For example, the traditional Simultaneous Localization and Mapping (SLAM) is a technique used in an outdoor environment that allows to obtain a representation of the environment, but it is not regularly used in underwater environments because these environments are difficult to study, namely because:

1. The size of the map and the computational cost are high because many calculations need to be performed;
2. There are many limitations in using different perceptual sensors because there is a lot of noise, as well as the presence of microorganisms;
3. Data association is a complex task, especially in cases where there are a large number of features.

As far as future prospects are concerned, there are 3 paths that this technology is taking [16]. The first is cooperative multicomputer navigation and localisation, which solves many problems as this approach coordinates navigation by sharing sensor data. Next is SLAM based on random finite sets, which allows for more accurate modelling and avoids the problem of data mapping, and finally SLAM based on deep learning, which is used to directly learn high-level features and employs machine learning and pose estimation to produce a map that contains both metric and semantic information, i.e. it presents geometric estimates and the class of all or part of the objects represented in the scene. It is important to point out that semantic information can enable a robot to better discriminate locations and perform higher level tasks, such as object manipulation. However, semantic SLAM is a more challenging task than traditional recognition, as the difficulties involved coincide with the complexity of integrating semantic information.

3 STATE OF THE ART

In order to recognise the limitations and difficulties of using some sensors and techniques in the underwater world and to improve our understanding of this environment, it is important to research and analyse the current state of the art.

The chapter is organised as follows: section 3.1 gives an overview of the state of the art of approaches that have made it possible to reconstruct the scene in which the robot navigates. Then, section 3.2 presents some work in the field of object recognition in water, and section 3.3 summarises some advances in semantic mapping that allow the robot to reconstruct the environment while identifying the objects present in it. Finally, section 3.4 describes the current state of the art in terms of available datasets, which helps in the development of some techniques to check the robustness and quality of the results.

3.1 UNDERWATER SCENE RECONSTRUCTION

Currently, many maritime surveillance tasks such as mine clearance, inspection and monitoring of structures or the seabed are carried out using data collected by AUV or Autonomous Surface Vehicles (ASV) with little or no human intervention. These vehicles are therefore able to operate in difficult underwater environments without direct human control and make some decisions autonomously during the mission. As technology advances, there are more and more applications for these vehicles that contribute to a better understanding of the oceans and the development of new solutions for various industries. Otherwise, the Remotely Operated Vehicle (ROV) is controlled to the water surface without posing a danger to humans. They could be used for data collection and access to remote locations without the need for human divers.

3D reconstruction is an important area in robotics because there is a lot of unknown information about this environment: the biodiversity to be analysed, corals to be observed, structures to be inspected, etc. These reconstructions are crucial to monitor the condition of structures, navigate a vehicle or manipulate objects. For these reasons, it is important that the final map is of good quality and that a wide range of information can be extracted. The data can be captured by one or more vehicles, which speeds up acquisition but can make reconstruction more difficult and computationally intensive. As mentioned earlier,

the sensors that can collect data to obtain a useful 3D reconstruction underwater are the acoustic and visual sensors. Therefore, in [17, 18] a preliminary 3D mapping in an underwater scenario based on the registration of six degrees of freedom (similar to the registration required for image stitching) was presented. The Tritech Eclipse sonar, a MBS with a beamwidth of 120°, was used. As the data is noisy, it was first necessary to perform a pre-filtering. To achieve the reconstruction using the proposed method, three main steps were performed: plane extraction, pose registration by plane matching and finally polygonisation. In this way, an area of 110×70 m could be reconstructed with only 18 planes of a river gate. This work showed that the registration of planes recorded with sonar without Inertial Navigation System (INS) or position sensor provides good results. Later, a new approach to create a 3D reconstruction of the seafloor using a MBS, a Doppler Velocity Log (DVL) and a Inertial Measurement Unit (IMU) was presented [19]. Thus, the described method allows the integration of localisation with navigation control for coverage and local control for tracking to cover an area of the seabed with a AUV in a previously unknown environment. In this scientific work, there are four main systems that are critical to the success of the reconstruction: localisation to calculate the position of the robot (using the DVL sensor); vehicle dynamics to estimate the speed of the vehicle; obstacle detection system and specific task-oriented sensors to obtain the data that allows the creation of the map. For this purpose, a downward-facing multi-beam sensor is used to collect data about the missions. To ensure a correct 3D reconstruction of the terrain, it is also important that the vehicle collects the data using an efficient algorithm for area-wide path planning. The reconstruction takes place after the data has been collected, i.e. offline. In the same year (2016), a 3D reconstruction method was developed that utilises data from a 2D MBS [20]. This sensor provides high-resolution images even in highly turbid water, but there are some limitations to their operation, namely the reflectivity of the materials or the need to place the echosounder deep enough (to ensure that it is free of disturbances that cause reflections). To create a 3D map with this sensor, the third component of the data (elevation) is missing, which causes some difficulties in interpreting the data. To solve this problem, one can use the relationship between the sonar data and the topography of the scene (which is determined by the reflection intensity and the shadow length). It is important to point out that this step is the most challenging in this method. The shadows also make it possible to determine the volume and edge boundaries of the objects, which could be important. This method is suitable for underwater environments and for real-time applications to include this method in obstacle avoidance tasks. The authors have also shown initial results

of merging texture information provided by a visual system with the sonar data to obtain a more reliable and intuitive 3D map.

The maintenance of structures or offshore installations is of vital importance and the monitoring, inspection and detection of faults is a current area of research. Recently, two different methods have been proposed in [21] to obtain 3D reconstructions of objects within short distances and in all weather conditions with a large aperture IS. This type of sensors is more suitable for short distances, as the images can be blurred and of lower quality with increasing distance, making the final reconstruction more difficult. The quantitative results presented show a good performance in terms of error (there are fewer outliers), but it always increases that the error of the motion estimation drifts. It is also important to point out that the sensor is less often affected by the multipath effect thanks to its smaller aperture. This method allows the scalability of the sampling (it requires less memory), which is an important aspect for underwater vehicles with limited memory resources on board. The compared methods are based on blind deconvolution with a spatially varying kernel and non-linear carving. Deconvolution provides more detailed maps, but is not iterative. Carving requires more computational memory, but can build the map iteratively. The fusion of different sensor types is a solution to many problems, but has some limitations. The authors believe that the method needs to be improved in the future so that the vehicle can perform different and more constrained trajectories on the sensor data, with curves and lateral movements. An AUV capable of navigating safely in complex, physically constrained environments is a crucial tool for cases that are dangerous to humans, such as cave inspection, which is another complex application of 3D reconstruction. For this reason, it is crucial to improve the accuracy of localisation modules in order to perform navigation and mapping tasks. Bringing a robot into a cave is a difficult task, mainly because of the configuration of the robot platform and the required on-board systems. Therefore, an approach for 3D reconstruction of caves was presented in [22]. This is an important task as these structures are complex, confined and restrictive environments. For this purpose, a robot with 3 main types of sensors was used: navigation, acoustic sensors for remote sensing and mapping and visual sensors for ground truth verification. Although the visual sensors are a common choice for AUV missions, their range is limited to less than 10m, as mentioned earlier. However, they are a good choice for checking the trajectory and ensuring that the vehicle maintains a constant distance from the ground. A MSIS was chosen because it has a 360° Field-of-View (FOV), but it is important to point out that it has a low resolution and a slowly rotating mechanical head. A problem with this sensor is the fact that the reflected echo could have originated anywhere along the corresponding height arc -

35°. In this work, two Tritech Micron sonars are used simultaneously, one in horizontal and one in vertical configuration. One provides planar views of the environment and is used in a SLAM algorithm to determine the relative displacement of the vehicle. The other enables the capture of the 3D shape of the cave with a resolution of up to 15mm. Both pieces of information are fused and the results show that the quality of the map depends on the quality of the trajectory estimation, the resolution of the vertical sonar and the speed of the vehicle. The results are visually appealing and allow the structure of the cave to be visualised by projecting the extracted points from the sonar over the estimated trajectory from the SLAM algorithm.

More recently, the use of simpler sensors such as MSIS for reconstruction has increased, but is usually treated as a two-dimensional problem. Thus, a method has been proposed to reconstruct the shape of various submerged objects, and it has been proved that the ambiguity of this sensor can be solved by scanning the same object from multiple viewpoints [23]. Thus, probabilistic occupancy estimation is used with a new method that enables the inaccurate MSIS measurements to be minimised by efficient data structures. Another way to obtain 3D information is to use two sonar devices. In [24], a Forward-Scanning Sonar (FSS) and a Profiling Sonar (PS) are proposed to work together to provide a complete mapping of the observed scene. The data from one sonar device can be transformed into the data from another device using geometric constraints. The main strategy is divided into three steps: carving out the area without object with an occupancy grid, then predicting the elevation angle of the FSS using the PS measurement, and finally performing the particle filter using the generated particles by calculating the weight and the acquired data of the FSS.

Another important task today is the problem of SLAM in turbid water. For this reason, a novel approach was presented in which the main source of information is a Tritech sonar and the ROV was additionally equipped with a depth sensor that provides an estimate of the 3D position of the vehicle [25]. In this paper, a multi-stage pipeline was proposed for localisation and mapping in shallow waters with currents using a robot without inertial measurements. Thus, this method uses a feature-based incremental smoothing and mapping solution, which includes three main steps: filtering some erroneous measurements, applying the SLAM algorithm, and performing accurate and robust Gaussian occupancy mapping.

The visual systems can also be used to generate 3D representations of the scene in underwater environments, provided the environmental conditions allow it. As their use in this environment is often very difficult, there is less work where the visual system is the only source of information. In [26], a method of 3D reconstruction is presented

where it is performed by a pair of cameras based on the triangulation of the common features between the left and right images. This type of approach often assumes that the scene is flat, which is not always the case and can lead to inconsistencies in the final result. In this work, this assumption is not taken into account and the results do not show any distortions. By using a SLAM technique with loop detection to estimate the successive camera positions, a texture map and a depth map of the scene are created. The method showed good results with sufficient resolution to enable the detection of underwater structures. However, this map was processed offline and required about two to three times the recording time (in a study with two hours of recording time, the map is ready after four to six hours of processing).

Apart from the importance of improving the accuracy of localisation modules in constrained environments such as caves, it is important to point out that there is still much to learn about them and their volumetric models to get a better overview of their size, shape, connectivity, etc. To understand the volume of these caves, it is important to know the volume of flow through the channel system. Nowadays, this work is limited, namely by the use of visual sensors, as these are limited in dark environments. Therefore, an approach utilising the presence of artificial lightning was proposed, as caves are an extremely dark scenario [27]. In this way, it was possible to create a model of a traversed area (based on the boundaries). Thus, a video light cone is used to identify the walls by a pair of cameras (the movement is also estimated by the cameras). Artificial lighting is necessary in many cases, but leads to some undesirable effects if used incorrectly. The developed approach makes it possible to obtain a volumetric map and the camera trajectory by combining successive images and relying on an advanced visual odometry algorithm. At the same time, it is possible to derive the scale from the stereo reconstruction.

Another strategy that has been used in recent years to obtain a useful reconstruction is the fusion of perceptual information from different sensors, namely visual and acoustic, whenever visibility conditions allow. Given the difficulties of cameras in the underwater environment and the fact that sonar can operate in a wider field of view, this has been a promising area of research in recent years. In this way, better results can be achieved in terms of the quality of the final reconstruction. However, this task can be difficult as two different pieces of information with different resolution, speed, etc. need to be combined. One of the most important problems in data fusion is the calibration of the sensors. In [28] suitable methods for this task were proposed using only 5 correspondences between the two sensor fields of view. In addition, a new 3D reconstruction method was proposed in which the Gaussian model is applied to the uncertainties in the rectangular image positions. These uncertainties are caused by

the conversion of the sonar measurements into rectangular images. The main difficulty in applying these methods in real time is that an automatic robust and accurate matching of the corresponding features in the optical and acoustic images is required. For this reason, and because the main objective of the work is to evaluate the calibration and reconstruction methods, the matching of the features is performed manually. The calibration methods show results that are accurate and effective. The experiments are carried out in a basin. For this reason, the sonar images were distorted by several echoes from the water surface and the walls of the basin. The reconstruction is done by triangulating the matching points in the visual images and sonar data. The approach is promising and shows potential when applied under different environmental conditions. However, the solution of the correspondence problem is performed manually in this work. Later, a system consisting of a camera and a multibeam echosounder mounted in a rigid frame so that there is an overlap area between the provided information was presented[29] . The success of the calibration process of both sensors in the same coordinate system is crucial for the correct fusion of the information. This is a difficult task as the sensors have different resolutions and the acoustic data is noisy. For these reasons, the authors considered it the best solution not to rely on correspondences. Therefore, a method for calibrating a system with a camera and a laser rangefinder was adapted. The map obtained allows the simultaneous verification of more complete information: the points of interest, the specific depth and the reflectance value.

Another way of representing the environment is the mosaic technique, which makes it possible to combine different images with an overlapping area in order to visualise a larger area than if we were looking at a single image. This technique has mainly been studied with images provided by visual systems. Over time, good results have been obtained, especially with underwater visual imagery, which presents many processing challenges to obtain visually appealing maps with the main goal of better observing the scene [30]. Some works with overviews summarising the main steps and their problems in creating larger complete mosaics from visual images have been presented recently, such as in [31, 32]. In addition, some challenges and strategies to solve the different steps are presented, compared and the main points discussed. This type of work is crucial, especially when it comes to applying these concepts to underwater data and often using different sensors. Since visual sensors are somewhat limiting for activities that require a certain maximum distance between sensor and target (usually 2 metres), attempts are being made to use acoustic sensors for this task. With FLS, for example, it is possible to obtain useful information about the structure of the environment, e.g. at short distances like the cameras, but in dark environments when the visual

data is unsuitable for interpretation. Therefore, video processing of sonar images is increasingly used for perceptual tasks, especially for the problem of image registration. In [33] it was shown the importance of applying different pre-processing techniques to improve the quality of the data, i.e. to correct the geometric distortions and make some radiometric adjustments to each image. In image registration for forward-looking sonars, i.e. with acoustic imaging sonars, three types of methods were distinguished: point extraction, region-based method and Fourier domain translation and rotation calculation [34]. In [35] a comparison between the latter two registration methods was performed and it was proved that for the tested datasets the Fourier approach performs better and ensures the accuracy of the final results. However, since the main interest in this work is to apply the same assumptions as for the visual methods, which are better studied and researched, the focus was put on the point extraction category. The different geometry and the fact that these acoustic images are in a polar coordinate system can make feature extraction and matching difficult, but some extractors such as Harris Corner can be used [36]. With FLS images it is possible to construct representations over the scene, with a particular interest in localising good quality targets in the final mosaic. In [37], the seabed was assumed to be a flat surface, and the process involves the key steps of feature matching compression to compute the transformation between overlapping images, this matching process, and some blending approaches to create a seamless mosaic. Another complexity investigated by the authors was the dynamics that can often occur in this environment. It was also shown that there are many variations in the appearance of features and that this is often the most difficult step of the process related to movement during acquisition, which can lead to inaccuracies in registration. In 2016, it was shown in [34] that it is possible to register a sequence of acoustic images based on tracked features in terms of intensity, texture and shape. A particle filtering technique was used for this tracking, and motion estimation is achieved by fusing different types of features. This method has been tested in unstructured environments where the extraction of conventional features such as points, lines and curves is difficult. It has also been shown that texture and shape features are more credible than intensity distribution features. However, only a single ROI is tracked until it disappears from the FOV of AUV. Often there is a need to investigate this type of application for a featureless scenario, and in these cases feature extraction is very difficult. Therefore, a work has been presented in [38] whose aim is to create an acoustic mosaic that can be used in target detection, especially in mine action. It is a real-time application, which is a good feature and, according to the authors, allows the mapping of any area with any dimension. To solve the problem of registration in featureless environments, the Global

Positioning System (GPS) system is used to calculate the displacement between two acoustic images and some correlation techniques are applied to correct the drift that may occur. Later, the same authors improve the real-time mosaicking to apply it to large areas, so that the method can be used whenever the optical cameras are not [39]. This makes it possible to map both large and small areas while maintaining the real-time function, i.e. without having to wait for global alignment or other optimisation processes, which in most cases take a lot of time.

3.2 OBJECT RECOGNITION

Object classification plays a prominent role in robotics as it is crucial for various tasks such as manipulation, obstacle avoidance, etc., but it is even less proven in the underwater world than in the air. Marine researchers therefore need automatic image analysis tools to easily recognise objects. To this end, much research has already been done for real-world applications. As described in the previous chapter, object classification follows identification. This means that first the existence of the object is established and then a classification is made to characterise the identified object.

There are a variety of solutions for object identification of objects with visual sensors and consequently with image processing techniques in a conventional way. A comparative analysis with some of them, such as frame differencing, optical flow, temporal differencing or background subtraction has been presented in [10]. It has been shown that the most commonly used method is background subtraction, which has a better balance between accuracy and time efficiency. The most important advantages and disadvantages of the individual methods are listed in Table 2.

Table 2: Advantages and disadvantages of the various traditional methods of object recognition.

Method	Advantages	Disadvantages
Frame differencing	Adaptation to dynamic scenarios	Low accuracy with dynamic objects
Optical flow	Accuracy when objects are moving	Many calculations Sensitivity to noise
Temporal differencing	Adaptable to dynamic scenarios	Error if the object is moving and stops
Background subtraction	Consistent in terms of results	Errors with background movements High memory consumption

To determine whether image processing-based algorithms are promising for this type of application, it is crucial to investigate them, their properties and limitations, and to perform analyses with direct comparisons using the available datasets. For example, a novel technique for finding artificial objects was presented that improves on another

work previously carried out by the same authors [40]. This approach was tested with different datasets available online, which differ in terms of camera quality, depth, lighting conditions, etc. In a first step, it was therefore necessary to extract the regions that stand out from the background, i.e. the Region of Interest. These ROIs are then narrowed down in two main steps to facilitate classification: Image segmentation and subsequent contour shape validation. Pose estimation can be performed with monocular data if the object is known and has a regular shape, or in contrast with stereo data. The point clouds obtained from the stereo system are often not dense enough, which affects the accuracy of pose estimation as it depends on the map disparity. The results of this work have proven that the method is robust to the disparity and achieves good precision in real underwater scenarios. If the intended object can be clearly determined by the shape of its boundaries, one possibility for object detection is to use keypoints or small parts of the boundaries [41]. However, these boundaries are often noisy, interrupted and blurred. In these cases, it is important to emphasise the boundaries using gradient maps and retain only the most contrasting points that represent the boundaries. The main advantage of this approach is the low complexity of the whole process. In 2017, a novel method combining the use of global colour contrast, intensity and transmission information was proposed to achieve better results in object recognition with only one camera [6]. So in the first phase, the features (intensity, colour and transmission) are extracted from the raw images (without any restoration or enhancement to achieve higher efficiency). Next, the colour contrast is increased, the ROI was identified and the global contrast of some features is calculated and combined. The extracted ROI is filtered and the object is finally segmented. However, this work assumes that the object is present on the image. This means that the method is not able to react correctly to cases in which there is no object on the image. Compared to other approaches in the literature, this method provides better results in terms of visual quality and eliminates noise in the background.

In many cases, the techniques that can be used for visual images can also be applied to sonar image data. For example, in 2012 [5] presented a method for underwater object detection for multi-beam FLS. This work attempted to address the two limitations mentioned above and also presented some improvements in terms of efficiency and speed, which are crucial in these applications. For example, the method calculates the features quickly and reduces the load as it works with smaller image sections. This made it possible to apply the method in real time on board an autonomous vehicle. The method takes into account environmental features and other geometric knowledge about the problem area and therefore does not require training data. The

method achieves good results when tested on a robotic platform and in a real environment.

When we want to evolve and separate two classes, for example, we often have to resort to machine learning techniques, because the conventional method is not able to learn the most important features of each class. In 2007, a method was presented that enables the classification of objects based on multi-field features and Support Vector Machines (SVM) [42]. The SVM is usually used in cases where two separate classes need to be classified, as in the proposed context. It is worth noting that the process of feature extraction from the data of a non-imaging acoustic sensor is very different from that of feature extraction from an image. This feature extraction is a crucial step and must be effective for the classification process to be successful. Therefore, four types of transformation can be used to extract features of the object from the acoustic signals. Using different feature extraction methods can be helpful to distinguish targets in reverberation, which is one of the biggest challenges when analysing acoustic data.

In this way, the features were divided into four categories:

- Coefficients of the Auto Regressive model in the time domain: these are features based on a time domain that reflects the response to the vibrations added to a target;
- Features of the power spectrum: this reflects the frequency distribution so that the amplitude statistics and shape of the spectrum can be extracted;
- Features of the Wigner-Ville distribution: the features are extracted simultaneously on the basis of time and frequency
- Features of the bi-spectrum: they are extracted based on the higher order statistical feature

Finally, the different types of features can be combined in one feature vector and so it is possible to apply the main classification. The SVM was used in this work because it promises to solve the classification problem with a small sample. This method was compared with two traditional classifiers: Gaussian and *K*-nearest. The results show that the developed approach can effectively distinguish between two classes and is therefore more suitable than the other two classifiers.

When it comes to classifying a set of object classes, it is often necessary to apply a more complex technique. There are some conventional supervised learning techniques for image classification, such as Maximum Likelihood (ML), K-Nearest Neighbour (KNN) and SVM [43]. These techniques require a training dataset to teach the classifier to define the decision boundary. Although these techniques require a lot of time for the training phase, the errors can be easily identified

and corrected. The SVM classifier optimises the use of training data as it assumes that there is no prior information on how to classify the data. It is also less prone to overfitting and requires less computational memory compared to the previous method. However, it is not suitable for large and noisy data datasets. The difficult factors in underwater scenes lead to non-linearity in the image, which makes it difficult for conventional supervised methods to adapt to the characteristics of the objects of interest in the images. In this context, data-driven classification models such as neural networks are more suitable. Deep learning has already proven to work well on large and diverse datasets, but suffers from overfitting when applied to small samples. Therefore, a method based on a Convolutional Network (ConvNet) whose architecture can learn invariant features was proposed [11]. The multiple layers that make up the ConvNet can enable the extraction of higher level features or the grouping of similar features in the same location or even the same feature nearby. Typically, these multilayer networks use supervised gradient backpropagation, but this requires many labelled samples. In this work, it has been shown how important it is that the ConvNet can be trained with very few labelled samples, as this reduces the effort and time required to label the data. This efficiency is achieved, for example, by using unsupervised methods in each phase. In this case, a Predicative Sparse Decomposition is used, i.e. an unsupervised learning method that trains a sparse encoding dictionary and an efficient predictor function that produces approximations to sparse representations.

The ocean today represents over 97% of the Earth's water, and to preserve it, critical marine habitats must be conserved [44]. Classification of fish species and plankton are applications where the impact of research has increased before 2015. The classification of corals is also an important task, but one that poses many challenges, mainly due to their diversity and their boundaries, which are often unclear and organic. In addition, the recognition of seagrasses has not yet been studied using deep learning methods, which can be a problem. Seagrasses are vital for the stabilisation of sediments or even as a food source for marine animals. Therefore, the effectiveness and robustness of learning methods can be crucial and represent a great opportunity for analysing the deep seabed.

There is a lot of work in the field of fish species recognition [45], but often the samples per data set are not very numerous. Therefore, a paper was presented in 2017 that provides good results with small sample data [46]. The method therefore starts by applying an improved median filter, which only processes the pixels affected by noise and presents more appealing images with higher quality, where no relevant and detailed information is lost. Next, the ConvNet is pre-trained with images from a large dataset so that the network becomes sensitive to

information such as colours, textures and edges. Thirdly, the network processes fish images to fine-tune the pre-trained neural network. The final method showed effective classification of underwater fish species with an accuracy of more than 80%, but can also be extended to other object classes. This result is very important as fish detection is one of the most difficult applications due to the movement of fish and species diversity, but also one of the most important for monitoring the status and trends of relative abundance, composition, size and biomass of fish stocks. A similar solution was later presented to compensate for the limited labelled data as well [47]. Therefore, cross-layer pooling using a pre-trained CNN with underwater video images is proposed. In this work, classification accuracies of 89% and 96.7% were achieved with the proposed dataset and a literature dataset, respectively, which is an incredible result for real-world contexts.

Due to the poor quality of images captured in the underwater environment, a new deep coding and decoding convolutional architecture was proposed to mitigate the effects of noise in the images, as it was assumed that the deconvolution operation would remove most of it [48] . The deep convolutional features provide higher robustness for the recognition task compared to other traditional extractors: Scale-Invariant Feature Transform (SIFT), for example. In the proposed method, transfer learning and data augmentation were additionally used. The deconvolution method was used for fine-grained feature extraction by refining the features after convolution. In an attempt to optimise the model, the authors apply transfer learning by training the model with a large dataset that facilitates the missing data. The learning parameters are then transferred to the network. When classifying fish species, the results show good accuracy in cases where the other methods cannot solve the classification. It has been shown that transferring more knowledge improves the results (increasing accuracy from 73% to 98%).

Acoustic sonars are also often used for harbour surveillance and help with important tasks such as avoiding obstacles or checking whether a diver is nearby. It is crucial that the can recognise what is seen on the acoustic image, and this is a challenging task, especially because the images have a low resolution. Nevertheless, a paper was presented in 2018 that deals with object recognition by the sensor, with the aim of automatically classifying 5 categories: Fish, Hull, Pile, Stone and Float [49] . There are 4 important steps in this method: image enhancement, segmentation (to separate the objects from the background), description of the image and finally the CNN classifier is applied to the images. The main motivation of the authors is to investigate whether transfer learning can help in cases where sonar images with lower resolution and smaller datasets are available. There are two main approaches to transfer learning: using a CNN as a

fixed feature extractor or relearning a CNN classifier to learn a new set of classes by training only a subset of the network layers. So in this case it would be possible to use a dataset like Imagenet for the training phase, but this dataset does not have much similarity with the images. For this reason, the features extracted from Residual Neural Network (ResNet) - 50 - were used, with the last layer removed. Finally, these features are applied in a SoftMax and a SVM method to evaluate the performance of the two approaches (CNN-Softmax and CNN-SVM). The results show that transfer learning can be an alternative to using manually created features and that CNN can provide good results even for low-resolution FLS images. In general, CNN-Softmax performs better than CNN-SVM. Otherwise, resorting to other strategies, such as fuzzy logic, can provide good precision in an object recognition method. Therefore, an adaptive Adaptive Neuro-Fuzzy Interference System (ANFIS) was presented, which is a hybrid system that facilitates the learning and adaptation of the developed methods [8]. For this purpose, both neural networks and fuzzy logic principles were integrated with the main goal of capturing the benefits of both in a single framework. The authors present a method in which ANFIS is trained with sonar images, the ANFIS applied to image tests to extract the objects, the regional features of each previously recognised object are determined and the object is identified based on some features (length, orientation, solidity, area, etc.). Thus, an accuracy of 85% was achieved in the recognition of objects and fish, which is an excellent result in a real environment. It is also important to emphasise that this method is suitable when only limited data is available.

The world is constantly changing and technology must follow the same trend to use vehicles and robots more and more autonomously to make different types of decisions. One of the tasks of the new robotic world is semantic segmentation, where the robot knows which pixels in an image belong to which classes. Semantic segmentation is an important task for the various fields of robotics, which often relies on visual data from cameras. Nowadays, it is often used in robots in underwater exploration, especially in autonomous navigation, when the robot needs to recognise the elements it encounters during the mission to avoid collisions, for example. It can also be used in object recognition and in the search for archaeological artefacts or even in the inspection of underwater structures such as platforms, cables or pipelines. Other applications can be found in marine biology and ecology in the identification of species, but also in the search and rescue of missing persons or in military defence through the classification of mines. For the semantic map to work well, the objects in the scene must be well classified and the image must be correctly segmented.

This task has its origins in digital image processing in combination with optimisation algorithms. The traditional methods initially considered a local view of the features in an image and focussed on local differences and gradients in the pixels. Later, adaptive thresholding, the Otsu algorithm and clustering algorithms were also proposed among the classical image processing methods. It is important to point out that clustering methods work well and can provide quite good segments in a short time, since the well-known k-means clustering considers all pixels and divides the image into k classes [50]. In this work, background segmentation of the objects was performed and the results showed that the algorithm can effectively segment the background of underwater images with low colour cast, low contrast and blurred edges. However, the traditional methods are generally not very transferable or robust, so the segmentation result of a single traditional method is poor in most cases. There are some issues with the accuracy of the results with traditional approaches as the data obtained in underwater environments have challenges such as colour distortion, low contrast, noise or uneven illumination and some important information is lost for these reasons.

However, in recent years, deep learning methods have emerged that allow the algorithm to return segmented images with some input, which is the main goal to learn and feed the system [51]. Therefore, these works need a set of training images to perceive the different visual features of each class and recognise exactly which pixels belong to that class. Therefore, most of the work is tested in outdoor environments, and when it is done for the underwater world, it is about one category, such as the classification and segmentation of fish or other animal species, or the classification of a few classes of objects. However, there is already some work that deals with this task in underwater scenes, and so an architecture with deep neural networks is proposed [52]. In this case consists of an encoder and a decoder. The pre-trained VGG-16 network is used as a feature extractor while the decoder learns to augment the lower resolution feature maps. The results show faster convergence and better accuracy than other methods in the literature. In addition, a neural network based image quality enhancement was developed and proved to be very efficient. One of the main difficulties in applying the different techniques in the underwater context is the availability of datasets, because in this context there are only a few datasets, and if there are, then only one class with different subcategories, such as fish, or only a few classes or configurations of perspectives between images.

The segmentation tasks based on training datasets often require a large number of samples per class with different perspective, size and appearance to achieve a good result when testing a model for semantic segmentation. In addition, the datasets for testing semantic

segmentation must provide a ground truth indicating which pixel belongs to which class. It is common that some of the available datasets for underwater scenarios that might be interesting for the semantic process do not provide this kind of information and therefore the data needs to be labelled, which in many cases is a manual task. As there are not many datasets with the best requirements for this task, many working in this area develop their own datasets but often do not make them available. Therefore, a real underwater dataset is created to compare the training with SegNet [53] and DeepLabv3 [54], which are considered the best deep learning segmentation architectures [55]. These networks are in most cases not tested in the wild underwater world and therefore this kind of work is very specialised. In this work, additional approaches such as fine-tuning and image restoration are used to verify that the performance is good. The dataset was put online, which is an important contribution as it was manually segmented, which is also a challenge. One of the most important factors influencing the final result in this task is the quality of the original underwater images, which is often very low. Nowadays, there are various approaches that are being tested, but complete solutions have also been presented [56]. Image enhancement is first performed based on a multiple spatial transformation, followed by a complex consecutive process to ensure effective learning about the image content. This ensures that the extracted rich contextual information significantly reduces information loss. Compared to other commonly used methods, the developed method shows high integrity of segmentation, high positional accuracy, high clarity of boundaries and high detail in subjective perception. In a detailed study, good results were shown for objects of different sizes (requiring a combination of global and local information), as the quality improvement benefited the accuracy achieved. There are still misjudgements and omissions for objects with low discriminative power and the authors believe that these results can still be improved. Although the results obtained in real contexts are not bad, it could be dangerous to work with them during a real deployment. According to the literature, it is possible to improve the results of semantic segmentation by adjusting the parameters of the model, resorting to some image enhancement techniques, increasing the number of images for training, using more balanced datasets, etc [57, 58, 59]. For several years, image quality was only improved with the intention of improving the quality of the results in terms of visual appeal to the user. However, the need to improve this quality is increasingly leading to better processing and higher accuracy of the results produced, reducing the difficulty of these more complex processing tasks.

3.3 SEMANTIC MAPPING

There are many situations in which it is important for a robot to combine reconstruction and object recognition in the same application. One example is the cleaning and inspection of chains, which requires efficient solutions. Therefore, [60] has presented a work that uses sonar technology as the main source of information. In this way, a scenario was used in which the vehicle navigates at a constant distance from the seabed (where a chain is located). The process is divided into three main steps: The detection of the chain link using sonar data and the calculation of the position based on the sonar origin. Next, the developed trajectory planner is used, which stores the successive positions of the chain links. Finally, the approach relies on the control system, which is responsible for guiding the robot to each waypoint in the sequence (keeping the orientation of the vehicle aligned with the last detected link). Thus, this work is a complete method that includes different techniques, but only the recognition of the chain and its mapping was analysed in the review of this work. For the recognition of chain links, two detectors were used for the data of FLS and MBS, using an enrichment technique followed by a pattern recognition based method. The map is created by registering the consecutive sonar images (vertical and horizontal) with the previously estimated trajectory. The map can be created not only with the sonar information but also with the visual stereo data. The experiments were conducted in a real scenario and demonstrate the accuracy in performing cleaning and inspection tasks.

Nowadays, it is becoming increasingly important for robots to understand their environment and also be able to navigate autonomously, create semantic maps and, in some cases, avoid obstacles. In this way, robots are able to understand the world better and better. There are already various applications in this context, but many of them only use two different techniques that allow the scenario to be mapped separately and the objects present in the scene to be segmented. Incidentally, most of these applications are intended for outdoor use. But again, it is crucial to know the behaviour of these approaches in the outdoor environment in order to better predict the behaviour in the underwater environment. Therefore, it was necessary to provide an overview of some related work that allows to better understand the problem and the challenges involved, as well as the main advantages when applied to the intended context and scenarios. It is therefore possible to combine a location categorisation based on a Bayesian deep learning method with a SLAM technique [15]. In this work, the main goal is to create semantic maps without specialised training to ensure that this approach is extensible and transferable, i.e. can learn new classes. Later, the same authors present a new work where the

focus is not on categorising locations, but on identifying the different entities of the environment of interest, also using geometric representations based on points or meshes. It uses deep learning for object recognition and traditional SLAM techniques to create meaningful maps [61]. With input data such as monocular images and odometry, a recently introduced implementation provides position estimates and the corresponding segmented object class [62]. Data association and loop closure are performed with semantic objects instead of 2D visual features extracted from the whole image. The objects are classified by implementing You Only Look Once (YOLO) [1] and the Oriented FAST and Rotated BRIEF (ORB) features are extracted in the area where the object is located. The accuracy in classifying the objects is 92.19%, which is an excellent result. Another work whose main goal was to perform semantic SLAM using commonly available flat surfaces as high-level semantic information was presented [63]. Thus, low-level Visual Odometry (VO)/Visual-Inertial Odometry (VIO) was combined with geometric information corresponding to flat surfaces extracted from recognised semantic objects. In this way, it is possible to create a sparse semantic map of the environment, which improves the metric estimates and has another advantage: No prior knowledge about the objects is required. The final approach is therefore fast, robust and lightweight and can deliver results in real time. Recently, in 2021, another approach was presented that aims to use a pose graph to represent the successive positions and integrate the 3D shape of the objects into a SLAM system by using semantic keypoints [64]. Although this approach involves higher computational costs, the final system is able to reliably and efficiently integrate semantic objects into a SLAM system, thus not only improving the performance of robot localisation but also creating an accurate semantic map. The method is able to work in real time, but at a higher cost and with a sparser representation.

Providing semantic mapping output is more challenging underwater than out, but in truth it solves many problems. For example, it can help automate underwater monitoring and inspection tasks that require a better understanding of the environment. Therefore, some work appears in this environment. One of the main differences in the application of semantic mapping in the underwater context as opposed to the outdoor context is the sensor used. Acoustic sensors are commonly used but have some characteristics that make this task difficult, namely the non-uniform resolution, the intensity variations of the individual bins, the acoustic reverberation, and so on. Therefore, it is often necessary that in the first steps an improvement of the image quality is made, then a segmentation that allows to obtain the objects, and finally the classification of the objects based on the

1 which is an object recognition model known for its speed and accuracy

hull area, the perimeter and the average and standard deviation of the intensity [65]. The classification process is based on SVM and two different approaches are tested: one based on 2D features and another based on a set of 10D features. It is observed that better results can be obtained with more features, proving that it is possible to automatically identify and classify objects in a real environment in a harbour with a forward-looking sonar. Later, again using a FLS as the main sensor, a new segmentation method is introduced which performs automatic local adjustment of segmentation parameters based on the average of bin intensities, and new statistical results are presented [66]. The use of this sensor has some disadvantages, because it is not possible to determine the vertical direction of the sensor, so there is an ambiguity problem. Sonar mapping often leads to dense point clouds that consume a lot of memory, contain a lot of noise and are only raw data without semantic information. Therefore, a method for semantic mapping was introduced where some artificial structures are known in advance [67]. The artificial structures complicate navigation because these structures require careful trajectory planning and there is a safety distance between the robot and the structures, which limits the resolution of the observations. The aim is therefore for the structures to appear on the MBS and be recognised automatically. In many semantic mapping applications, the main goal is to recognise objects and use them (or parts of them) as landmarks for SLAM. Thus, a 3D point cloud without colour information can be extracted with a fast laser scanner and a global 3D descriptor (after the segmentation process) can be used to identify and locate the objects, but in this kind of work only for objects that are already known [68]. Simultaneously with the object recognition, SLAM is performed to create the map of the robot's environment based on the positions of the key landmarks. This method can be used with the main goal of navigating the AUV in the vicinity of artificial structures and improving the quality of navigation [69]. However, it is necessary to use a laser scanner with a high data acquisition rate.

3.4 AVAILABLE DATASETS

Another crucial aspect of underwater perception research is the datasets that make it possible to test the methods developed in accordance with the objective. This requires that each dataset used covers the various aspects necessary for testing, such as sensor configurations, environmental conditions, the presence of ground truth, etc. However, in the underwater environment, there are not a large number of datasets with perceptual information, as this collection is not easy and is an expensive task that requires appropriate materials and communication

platforms and is often dangerous for divers. Furthermore, this task involves several problems with visual sensors, such as light absorption, turbidity and limited range, which reduces the number of good quality images. Depending on the topic to be researched and tested, the literature is reviewed to determine which datasets are best suited for the intended application and, if necessary, create one (if there is no other solution) that allows the evaluation. One of the most important aspects of object classification is the amount of data required to classify objects. This is usually high and the increasing search for images to test and experiment with some methods allows the creation of datasets of ever-increasing complexity, size and variety.

In this way, an extensive set of images covering different categories was created in 2009 by [70]. This was a real challenge for the authors and an important contribution to the scientific community. The main features of this set are: the large size (originally the set contained over 5247 categories and now consists of about 21000), the organisational hierarchy of the classes, the accuracy and the diversity of the images (position, size, viewpoint, etc.). Over time, the quantity of images and the number of categorised classes increase and the quality of the data improves. It is important to mention that the ImageNet also allows training without overfitting the data as there is a large variety.

As far as a better perception and understanding of the underwater world is concerned, the available datasets are scarce and when they are, they are often sparsely specified or do not cover the largest real-world conditions that are often necessary to achieve good results, for example in the classification of objects. Visual images are often used in this area as they can provide a large amount of information. In this environment, there are a variety of objects that need to be identified, e.g. critical human objects such as anchors, pipelines, mines, etc. or elements that may occur during robotic operations such as fish, other robots, divers, corals, etc. In this way, the datasets related to the biodiversity of fish or corals are the most frequently presented in the literature. In [71] a worldwide open repository is presented that allows the upload of various images of different corals to study their diversity. Later, in [72], some of the same authors present another dataset that can be used to categorise corals. The ground truth of this categorisation can be used, for example, in classification tasks. For more complex tasks, such as semantic segmentation of image content, datasets such as CoralSeg [73] can be used, where the different categories are labelled in the ground truth and can be used to train specialised networks. When studying fish species in detail, classification is sometimes crucial to better understand the habitat of these species and to recognise the fish under different aspects and conditions. Therefore, in [74] images are collected and provided with the main goal of fish classification or other important studies such as fish counting. In addition, more

complex tasks, namely automatic size estimation, need to be explored as overfishing is a problem in all oceans. With this in mind, a dataset is presented in [75] that contains data to assess areas such as fish instance segmentation, classification and size estimation. Both types of datasets are used for more detailed studies on different specific topics, but in this particular case they can only be a part of the final datasets that bring together different types of objects in the underwater world. Recently, another study has been launched to classify and identify litter found in the oceans, such as the Marine Underwater Environment Database (MUED) dataset [76] and TrashCan dataset [77]. Both datasets were created with the aim of object recognition or classification, so the ground truth is not the labelling of all pixels of each image. However, this is a part that is not useful in this particular context, as it is often not necessary to change the behaviour of the robot during the mission. If the algorithms are not only aimed at classification but also at segmentation of objects in an image, they need to be increasingly better trained, as already mentioned, since it is a dense classification that requires many examples with different perspectives and sizes for each object to be learnt. In this domain, the number of available datasets is smaller and many of them do not contain ground truth, which is a difficult task when they need to be fully labelled pixel by pixel in a manual task. This fact is one of the biggest obstacles when using some of the few datasets. For example, the UFO-120 dataset [78] is a general dataset that represents many classes such as fish, robots and plants. It was created with the aim of recognising the saliency region, which could be used for segmentation, but the ground truth only indicates that objects are present in each pixel, but not the class of the object, see Figure 3.



Figure 3: UFO-120 dataset: real images and the corresponding ground truth about the existence of the objects. Adapted from [78].

Furthermore, in most cases each image represents only one object, which limits the learning process. Often these datasets are also sparse in terms of perspectives for each class, poor in terms of diversity of classes, or contain only a few samples for each class, making it difficult to achieve good visual segmentation results. However, in order to use an autonomous vehicle to monitor an unknown area, the robot even needs to create a semantic map that provides it with a more comprehensive representation of the environment. For this purpose, it

is necessary to use a general dataset with different objects. Therefore, the Segmentation of Underwater IMagery (SUIM) dataset proves to be the most attractive for this task, as it contains a large number of classes that may frequently occur in the intended environment [79], such as human divers, robots, plants, wrecks or fish, see Figure 4.

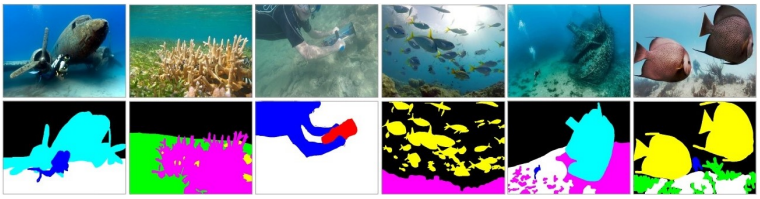


Figure 4: SUIM dataset: real images and the corresponding ground truth masks with the labelled classes. Adapted by [79].

As mentioned above, in some cases it is useful to use acoustic images to explore and test the concepts commonly used in visual images, such in 2D reconstruction of scenes. However, this task is even more complicated as this type of dataset is scarce and in many cases does not even exist. Therefore, it is not easy to test the developed algorithms as the data has to be collected after the experiments and in many cases the sensor is not yet available. *SoundMetrics*[2] has done good work in this area in recent years. The company develops and manufactures imaging sonars that can be deployed from a ship, a ROV or in an AUV. These sensors are used in many fields including military, oil and gas, underwater construction and law enforcement. The company markets 2 models of sensors (ARIS and DIDSON) and shows some examples of small surveys conducted with these sensors on its website . This is very helpful in selecting these sensors and at the same time it is possible to use some of them for small tests with real data. Therefore, they were asked to provide some trials and the corresponding tools to convert the raw data into images. The selected data refer to some trials that are suitable for the context of the work, namely those representing for example mines, ship hulls or pontoon welding beams for inspection structures. Figure 5 shows some examples of acoustic images that are part of the provided trials.

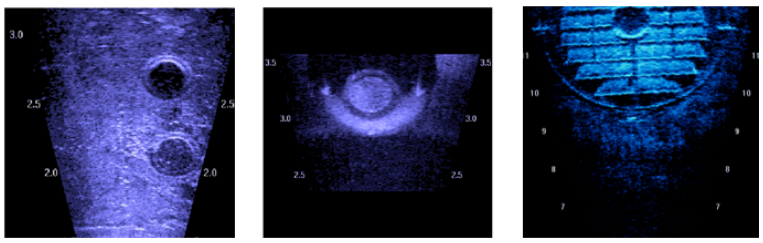


Figure 5: Some examples of acoustic images provided by *SoundMetrics*.

2 http://www.soundmetrics.com/ I would like to thank you in advance for providing the data and thus contributing to the developement of this work.

Part II

SCENE RECONSTRUCTION AND OBJECT RECOGNITION

4 TOOLS AND ROBOTIC PLATFORMS

This chapter presents the selection of tools and robotic platforms used to evaluate the methods developed for the underwater world according to the scenarios envisaged.

The chapter is organised as follows: section 4.1 presents the advantages and benefits of resorting to simulations to test the developed as well as the good selection of sensors to be used in underwater vehicles. Section 4.2 describes the simulation tool that was used to test some of the methods developed for a survey mission and the description of the datasets collected with this tool. Next, section 4.3 compares the sensors that are best suited to perform good perceptual data acquisition that can be used for scene reconstruction and object recognition and describes the vehicle that was used in some of the survey missions. Section 4.4 describes the datasets collected with this real system that can be used to test some crucial steps of the methods used.

4.1 INTRODUCTION

Whenever a new method for using an autonomous robot is developed or adapted, it is important to test its performance and behaviour under conditions that simulate real-world conditions so that a robot can later be used in the real world with greater safety and confidence.

For underwater environments, there is currently a lack of suitable datasets for real-world conditions compared to other areas. This is because it is difficult to obtain high quality data in this environment and these campaigns are very time and resource consuming. At the same time, there are some limitations to the available datasets. For example, they do not contain all the situations needed for the tests, or the sensor configuration is different than expected. Therefore, it is often necessary to use simulators at the beginning of the comparison and test phases of each approach or for intermediate steps. These can replicate some conditions of the intended scenarios, namely some difficulties associated with the underwater environment, using different sensors depending on their availability in the particular simulator. Therefore, they help in the preliminary evaluation of the developed methods and highlight some limitations that may arise unexpectedly. In this way, it is also possible to give a certain dynamism to the exploration of the environment.

After the simulation, the behaviour of the methods must be verified in real contexts and with data from real sensors. It was therefore necessary to select the best sensors for the different applications and combine them into a perceptual system, which is one of the vehicle's most important systems. When selecting these sensors for underwater applications, the specific context of the research, the environmental conditions, the operation and the type of data required must be taken into account. Therefore, different sensors can be combined to get a better understanding of this environment. Data acquisition has become a key element to increase the quality and efficiency of any robotic application. This task allows the collection of information about the environment that can be used for some post-processing to draw conclusions about the conditions of the environment or for online processing during missions. In a phase where new technologies, vehicles or approaches are being tested for a specific application, data availability is also crucial as there are few online datasets applied to this environment. In real contexts and missions, the acquisition system should be able to capture the data information and send it to the process modules so that the vehicle can make decisions in real time.

4.2 STONEFISH SIMULATOR

The simulation of underwater scenarios is of crucial importance in the context of underwater research, especially when testing and developing methods. In order to simulate the intended scenarios, the Stonefish[1] simulator was therefore selected as the main simulation tool in accordance with the findings from the literature in this field [80]. This is a C++ library that combines physics with the ability to render an environment that incorporates advanced hydrodynamics and is based on the actual geometry of bodies so that various effects on the environment can be simulated with realistic performance, such as rendering the atmosphere, ocean and underwater environment. The effects of wavelength-dependent light absorption and scattering have also been taken into account, which is an improvement over other simulators. The simulator can be used to create stand-alone applications or integrated into a Robot Operating System (ROS) architecture.

With this simulator, you can simulate how the vehicle explores a specific area on a predefined route. The vehicle can be equipped with various perception sensors such as cameras and depth cameras, imaging sonars such as FLS or MSIS or even SSS. They make it possible to detect the structure of the environment and record data if required. In addition, the vehicle can be equipped with sensors that are important for its movement, such as IMU or GPS, as this simulator also allows

1 https://stonefish.readthedocs.io/en/latest/

the simulation of scenarios with surface vehicles. A virtual odometry sensor can also be added to obtain the ground truth for navigation, which is crucial for comparison with data from motion estimation applications. To create the simulated scenarios with this tool, it was necessary to create some seabeds simulating port or archaeological areas with the objects normally seen in these regions, depending on the preferences of the exploration. To define the intended scenario, the simulator allows this definition through a code that allows to create different simulation worlds with generated terrains, different bodies, etc. They are able to reproduce geometry files representing the bodies, which can be dynamic and therefore move in the scenario, and obstacles, which are static bodies. The objects can be made of different materials that define the physical properties. This can be advantageous in some tests, namely in studies with acoustic sensors, as their behaviour changes depending on the material of the object.

For the terrain, a set of points (mesh) can be defined in a flat grid with a structure in the z direction. Three maps are used to create this terrain: Albedo, Height, and Bump (also called Normal Map), which must match to get a good rendering for the scenario. These maps are created from the original image that represents the environment in question. The appearance of the scene is shown on an albedo map, a map that shows the inherent colour and reflectivity of surfaces. The second map, the height map, gives an idea of depth to simulate the three-dimensional details of a surface. Finally, the bump map represents the texture map.

Figure 6 shows an example of the three maps so that you can recognise the differences between them.

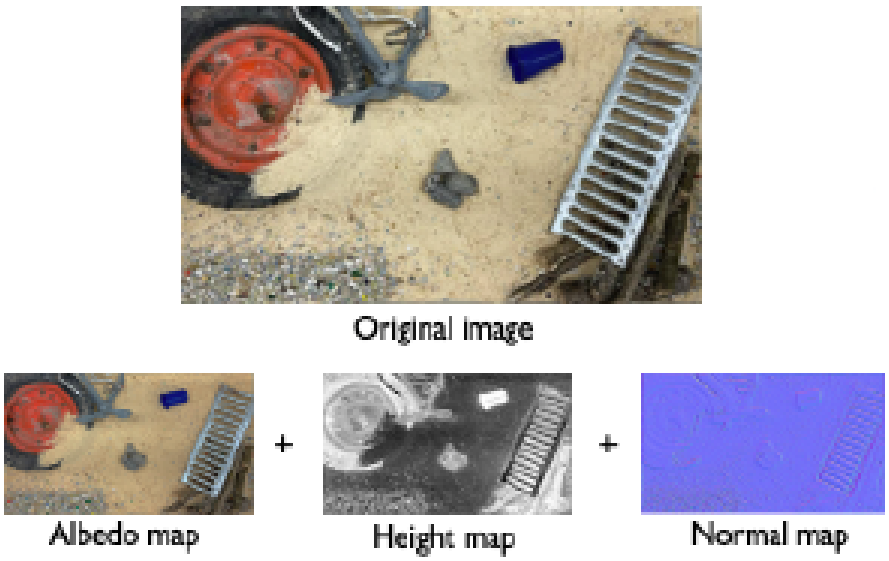


Figure 6: Example of the three maps used for the rendering process to create a terrain from the original image.

If you put the three maps together in the simulator, you will obtain a terrain that looks like a real texture. You can see the final result of the scenario in Figure 7.

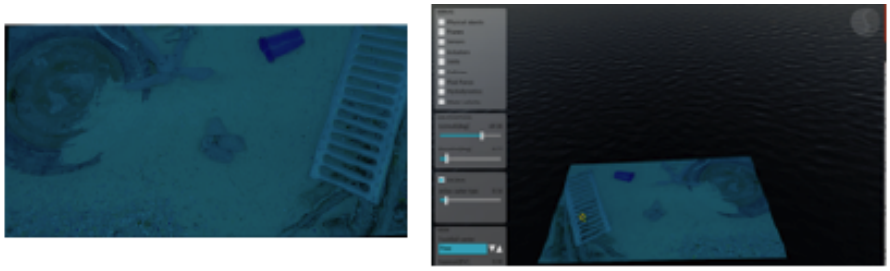


Figure 7: Example of the final result of the rendering process.

After the rendering process, the vehicles can be inserted into the scenario. For this purpose, the *Girona500*, an AUV provided by the simulator, had to be equipped with perception sensors to imitate a real future scenario. This vehicle is a simulation of the real version where it is reconfigurable and offers good performance in hydrodynamics with a compact size [81]. Therefore, a camera with the parameters of the Allied Vision Mako G-125B/C Gigabit Ethernet (GigE) camera was installed to replicate the real sensor. To simulate the acoustic data, a FLS was also simulated with the configuration parameters of the model chosen for the real applications (see the next section for more information). Figure 8 shows an example of the simulator and the data collected by both sensors.

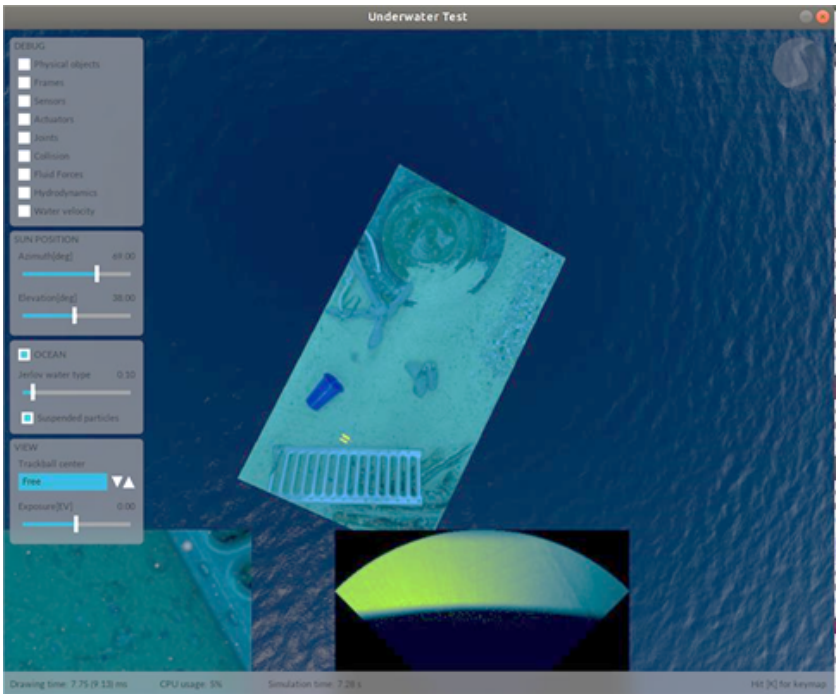


Figure 8: Example of the simultaneous acquisition of data by the camera and FLS in the simulator.

The grid can be seen in both the visual and acoustic images, although it is larger in the latter because the FOV of the acoustic sensor is greater. The acquisition rate of all sensors was configured to be the same, as the data should be compared later, if possible. In order to collect different datasets, it was also necessary to create different trajectories between different waypoints chosen by the user. Therefore, the robot can travel between two different points (point A to point B) with a simple strategy: First, the robot is turned towards point B, then it approaches this point at a user-specified speed and on a forward trajectory. When the robot is close to the desired point, the speed is reduced to a good approximation and the vehicle finds the final orientation again. Whenever the robot has reached the desired point, it checks whether there is another point on the route or whether this is the final position. If this is the case, the movement is terminated, otherwise the robot continues to the end point. In this way, it is possible to carry out different experiments in the simulated scenarios that allow some approaches to be better compare some approaches and make decisions in a more appropriate way, as well as increasing the performance and improving the results of the methods developed before testing them in real conditions.

In this case, the simulator can be interesting to test the method developed for acoustic sensors, namely the mosaic techniques that use acoustic images, and the available datasets for this type of application are scarce. For this purpose, some scenarios are created with different common elements of the underwater to simulate the areas to be inspected (see Figure 9). The vehicle must be configured to capture different trajectories in a harbour scenario with a FLS that provides data for the scene in front of the vehicle.



Figure 9: Example of the scenarios developed for the depiction of harbour facilities.

4.2.1 Build Datasets

- ***Port_Mosaic***

  In this dataset, the vehicle *Girona500* performs a trajectory in a scenario of harbour structures, for example to simulate an inspection around some structures such as a wall with a lenght of about 2m. The trajectory represents a situation in which a previously visited area is revisited, but with the same point of

view. The FLS sensor is configured as a top-down sensor and the data provided refers to the seabed after the corresponding configuration. The main configurations of the FLS were a range of up to 3 metres, a tilt angle of 35° and a horizontal FOV of 40° to capture really good information - see Figure 10 to better understand the FLS model for these configurations.

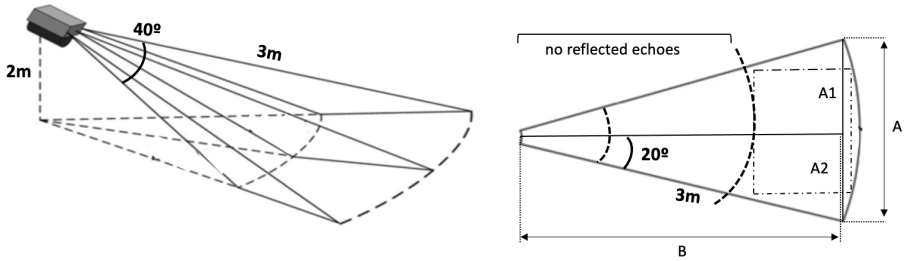


Figure 10: FLS model corresponds to the configurations during the acquisition of the dataset *Port_Mosaic*.

Therefore, the sonar data provided has 512 beams and 750 bins to simulate the real sensor. This is an acoustic image of 514×720 pixels covering a width of about 2m (A) and a height of 3m (B), i.e. an area of about 6 m^2. Figure 11 shows an example of the recorded acoustic images.

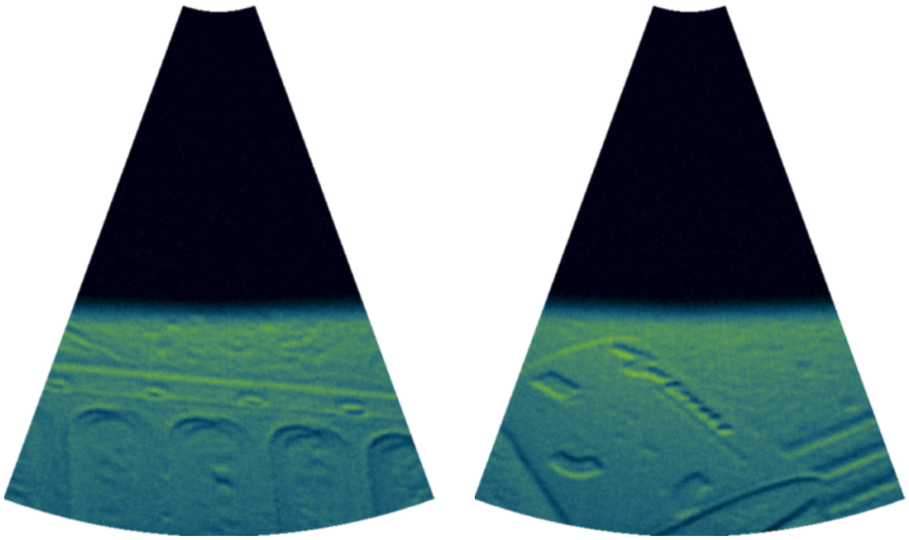


Figure 11: Examples of the acoustic FLS images obtained with the Stonefish simulator.

A total of 1067 images with a frame rate of about 7 fps were captured during the trajectory of the mission with the Stonefish simulator version 1.3.0. Figure 12 shows the trajectory performed by the AUV at a fixed depth z.

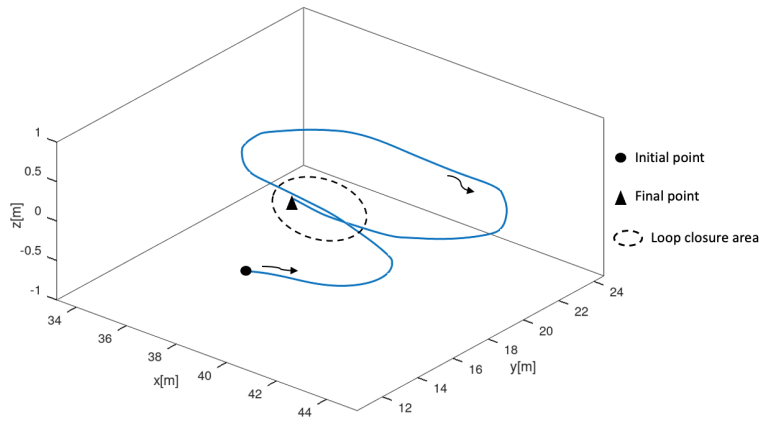


Figure 12: Trajectory of the vehicle during the *Port_Mosaic* dataset.

If you look at the images 10 and 11 again, you will notice that there are a number of pixels of the returned images of the Forward-Looking Sonar that are black and thus do not contain any valid and useful information. Since this dataset was created with the main goal of testing the mosaicking techniques with these images, it makes more sense to use only the useful part of the images. Based on this data, a Region of Interest is used for each image to discard the black area, i.e. without reflected echoes. Therefore, each image is cropped to a rectangular area to account for the effective visual information captured by the sensor. Based on the Open Computer Vision Library (OpenCV) C++ library, a ROI is therefore selected for each acoustic image to represent the effectively imaged area of the output. Thus, a bounding box of 250×350 pixels is used, which means that each FLS image covers an area of about 1.5 m^2, as shown in Figure 13.

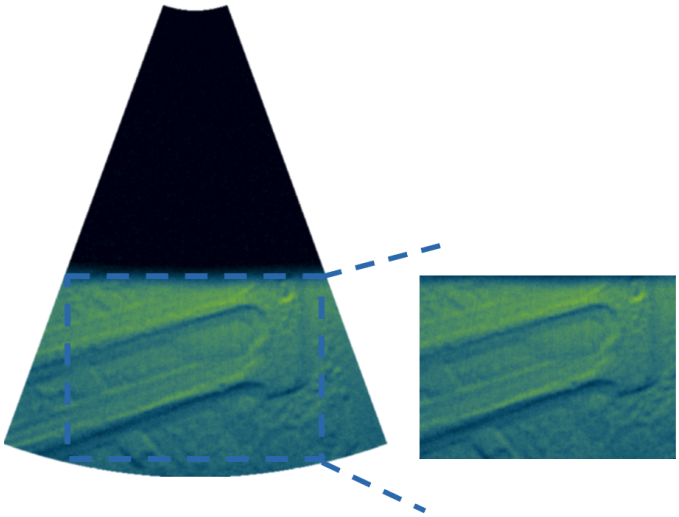


Figure 13: ROI that allows the selection of the part of the acoustic image that represents the returned echoes.

4.3 AQUISITION SYSTEM

Each robotic vehicle, whether autonomous or remote-controlled, has different modules with different goals, which vary depending on the purpose of the robot and the tasks to be performed, but many of them have common modules, such as those shown in Figure 14.

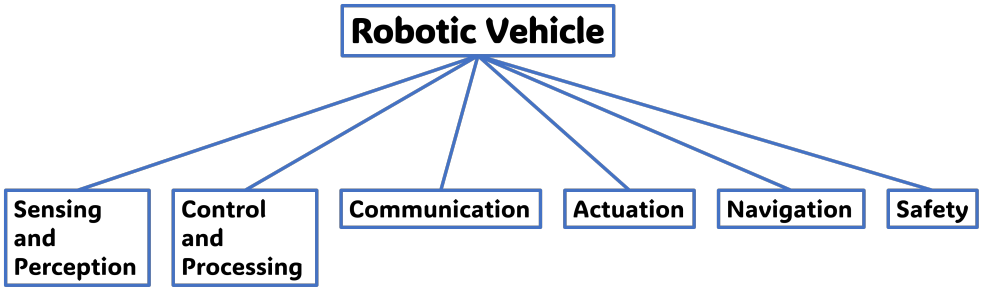


Figure 14: Main modules that are normally part of robotic vehicles.

They have various responsibilities during the mission, from controlling movement and precise navigation to collecting perception data. This module of data acquisition and processing has become a key element for increasing the quality and efficiency of many robotic applications and is therefore the driving force behind the studies and analyses conducted in this thesis. In order to achieve a good recognition and perception of the environment, it is necessary to first collect real data with quality. Therefore, it is possible to test and prove the assumptions and results of the preliminary tests with simulated data and available datasets. To do this, it is necessary to select the most suitable sensors and equip a vehicle with the main perceptual sensors capable of collecting data information that will allow the robot to better perceive the environment and even decide some actions according to the processing of this information. These systems refer to all the sensors of the robot that allow it to understand its environment. For this reason, it is crucial for the robot to interact with the underwater world and navigate in it. Otherwise, these systems enable the collection of data, which is often crucial to study and test certain methods or to make them available to the scientific community. In most cases, these are image sensors such as cameras or distance sensors that provide information about the distance between the robot and the objects. These sensors enable many tasks such as navigation and manipulation. To ensure the quality of performance of these techniques, it is often necessary to use efficient sensors and ensure that they provide high quality outputs and high frame rates.

It was therefore necessary to select new sensors and integrate others into the perception system in order to obtain the best possible system. All new sensors to be installed were evaluated and well planned to fulfil all the requirements of the perception approaches. Some

comparisons were made in terms of data provided, operating limits, consumption and dimensions.

Acoustic sensor

After some theoretical work, it was concluded that an imaging sonar device with a greater range had to be selected for the visual approaches. To this end, different models and their main characteristics were analysed. Sonar technology was initially proposed in the literature with the main objective of detecting icebergs, but there are now numerous applications for this sensor. In the military field, there are a large number of systems that detect, identify and locate submarines. It can also be used for acoustic mines and mine detection or for non-military applications such as fish detection, depth sounding and seabed mapping. So there are two types of sonar: passive sonar, which essentially listens for the sounds of ships, and active sonar, which emits sound pulses and listens for their echo.

As the main aim of this research is to improve the perception capability of vehicles, the choice of sonar technology fell on imaging sonars. By transmitting and receiving sound waves, these systems produce acoustic images that are similar to optical images and enable, for example, the monitoring of structures or fish species in dark and murky waters where optical cameras are of limited use. In recent years, imaging sonars (especially forward-looking high-frequency multibeam echosounders and side-scan sonars) have been widely used for environmental mapping and interpretation. These imaging sonars are efficient and effective tools that enable the identification of objects of interest by creating a sonar reflection or blockage that produces a sonar shadow. There are some imaging sonars, such as the scanning sonar, where only the illuminated area is visible to the user, or the rotating scanning sonar, where the visible area is magnified. The range is usually around 75 metres. Otherwise, there is another type of sensor, side scan sonar, which images both sides of the sensor, but the centre is incomplete and for this reason is not the best choice for applications that require a complete and higher level of understanding of the environment. Finally, forward-looking sonar is designed for surface vessel navigation and obstacle avoidance and images the forward area, but with coverage suitable for the intended application. It has a longer range, usually in the order of 200m. It is also important to point out that the forward-looking multibeam sonar has some other positive aspects compared to the side scan device, namely: the fact that the results are immediate because processing is required, and that they are more versatile in terms of configuration. However, the interpretation of the data can be more difficult and the total area captured is many times smaller because, depending on the configuration, part of the information returned consists of black pixels indicating regions

without echoes. Nevertheless, this sensor was selected as the most important perception sensor among the acoustic sensors. However, it was necessary to investigate the current state of the art in order to better understand its characteristics, advantages and limitations and to select the best model.

In practise, forward-looking sonar systems are primarily used for navigation in shallow waters, but they can also be useful for reducing the likelihood of dangerous disasters such as oil spills, reef groundings and whale accidents. Therefore, this sensor is a good alternative to optical sensing as it provides high quality data even in waters with high turbidity. However, there are often some limitations to its use that also apply to other acoustic systems and that the user must be aware of, such as:

- The resolution is inhomogeneous, as it is lower in the upper part of the image.
- There is an acoustic distortion caused by the movement during sonar operation. This often results in the objects being flat (making shape recognition difficult, especially for small objects)
- Loss of 3D information as it only provides data from two planes, leading to ambiguity issues
- Inconsistent intensity of pixels as it depends on water density, orientation of sonar in relation to objects, sensitivity of hydrophones, etc.
- Acoustic reverberation that creates multiple ghost objects.
- Acoustic shadows caused by other objects blocking the passage of waves.
- Noise caused by the reflection of the water surface, interference and engines of other vessels.

Nevertheless, this sensor has been used in recent years with good results in underwater scenarios, particularly in SLAM approaches, 3D reconstruction or even applications involving the detection and classification of objects or areas [82, 83]. In order to select the forward-looking sensor best suited to the intended context, it was necessary to make some theoretical comparisons and get an overview of the main features of the other sensors. Firstly, the main manufacturers of acoustic sensors were identified, i.e. those that offer sensors that are or could be used for forward-looking sonar devices. There are three major manufacturers: Teledyne, Tritech and Imagenex, which are well-known suppliers and have a good reputation in the scientific community.

- Tritech is still a leading company in the industry. The products in question are the Gemini segment of multibeam echosounders, the smallest in the world. Only two models are considered for the application in this system: Micron Gemini (labelled A.720im[2]) and the Gemini 720ik (labelled A.720ik[3]), see Figure 15. The main difference between the two models is the FOV and the resolution of the data obtained, which is higher in the second model. The other models have a higher operating frequency and can provide better high-resolution images at shorter distances, which is not the main objective in this context.



(a) (b)

Figure 15: Models of FLS from Tritech: Gemini 720im (a) and Gemini 720ik (b).

- Teledyne has two different segments of acoustic sensors that can be used as forward-looking sonar devices. The first is the Blue View P900-130, which provides high-quality sonar images in real time, but was discarded because it has a lower resolution than other models from this supplier.

  Next, there is the M-Series Mk2, a compact sonar unit with higher resolution and low power requirements, which is a crucial factor There are two different models in this segment (labelled B.900Mk2[4] and B.900-2250Mk2[5]), which differ in some features, but mainly in frequency, as the second model has both frequencies in the same sensor. The higher frequency is mainly used for close-range inspection as it works up to 10 metres away from the structures. The detection rate is higher than the previous models. A third segment of this provider is not considered as it is a solution for military applications with large dimensions, FOV and range. You can see the appearance of the two models in Figure 16.

2 https://www.tritech.co.uk/products/micron-gemini
3 https://www.tritech.co.uk/products/gemini-720ik
4 https://www.teledynemarine.com/brands/blueview/m900-mk2
5 https://www.teledynemarine.com/brands/blueview/m900-2250-130-mk2



Figure 16: Model of FLS selected from Teledyne.

- Within Imagenex there are many categories and different sensors, but in this context there is a multibeam category with information about image sensors. Two different models are presented that are similar and offer high speed and resolution. The main difference between them lies in the operating frequency and the FOV. The maximum range of the second model (labelled C.965A[6]) is higher (100m compared to 30m) and for this reason it is the first choice for comparison to select the most suitable sensor, see Figure 17.



Figure 17: Model of FLS selected from Imagenex.

Table 3 shows the most important differences between the sonar models of the three manufacturers selected for the preliminary comparison, in terms of operating frequency (Freq.), FOV, range, resolution (Resol.), power, depth operation and height in water.

Table 3: Comparison between the 5 selected models of FLS to analyse their main characteristics.

	Freq. (KHz)	FOV (º)	Range (m)	Resol. (cm)	Power (V)	Depth (m)	Height (kg)
A.720im	720	90	50	0.8	12-48	300/750	0.25
A.720ik		120	120	0.4	19-74	350	0.43
B.900Mk2	900	130	100	1.3	12-48	1000	1.2
B.900-2250Mk2	900/2250		100/10	1.3/0.6			1
C.965A	675	120	100	20	22-32	1000	0.045

6 https://imagenex.com/products/965a-multibeam-imaging

After making some theoretical comparisons and analysing the main features of the different models on the market, we have opted for the Gemini 720ik[7] model from Tritech, which offers the best price/performance ratio within the available budget. This model is a high-frequency multibeam echo sounder with a refresh rate of 97 Hz. This is an excellent price/performance ratio to produce high quality, video-like images and achieve the best continuous result in perceptual applications. It enables use in shallow waters, even with a maximum range of 120m. It offers a FOV of 120°, low power consumption and a small, reduced weight with an optimal range resolution of 4mm (the best compared to the other four models).

Visual sensor

Visual sensors are increasingly used in underwater applications, albeit with some limitations that affect their quality and complicate their use. This sensor makes it possible to better understand the vehicle's environment because it is a key to perception sensors, as they directly simulate human eyes. They provide a variety of information about the scene, including its appearance in terms of colour and texture.

The used model was selected because this sensor has already been used in various works and is characterised by its efficiency and the quality of the images provided. Therefore, an Allied Vision Mako G-125B/C[8] GigE camera with a 6mm lens was used in the perception system. The fact that its parameters are highly configurable and the high frame rate that can deliver images (30 fps) at full resolution (1292 × 964 pixels) is very interesting. It enables automatic correction of gain, exposure and white balance, which is an advantage when working in underwater environments with changing light.

Camera Calibration

The images are therefore captured and, due to distortions caused by the lens or by scattering or absorption effects, may have some distortions that need to be corrected. These corrections are very important, especially in cases where the pixel information must be used to obtain measurements such as distances, object dimensions or extraction features. For this purpose, a technique based on OpenCV was applied to several images of a chessboard taken underwater from different perspectives. The technique recognises the inner corners of the chessboard, as you can see in Figure 18, regardless of the perspective shown in the image.

7 The main features of this sensor can be found in https://www.tritech.co.uk/products/gemini-720ik.

8 The main features of the camera can be found in https://www.alliedvision.com/en/camera-selector/detail/mako/G-125



Figure 18: Some examples of chessboard images in which the inner corners are recognised regardless of the perspective.

After this extraction, the corners are used for the calibration process, where the intrinsic parameters of the camera, such as focal length, principal point and distortion coefficients of the lens, can be estimated. In this way, you obtain a vector and a matrix representing these parameters:

$$\text{Distortion Coefficients} = \mathrm{k}_1, k_2, p_1, p_2, k_3$$

$$\text{Camera Matrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix},$$

where f_x and f_y are the focal lengths along the$x$ and y axes respectively, and (c_x, c_y) is the principal point of the image.

Also using visual data, but for other applications, a system for structured light has been developed. This involves projecting a known pattern or structure of light onto a scene or object and analysing the deformations or distortions of this pattern using the images returned by the camera. In many cases, this process has been used to extract three-dimensional information from the scene or even to increase the quantity and quality of visual information in the images.

For this reason, an Osela solution was chosen for the projector. Lasers are available in many colours, with specific wavelengths available for each colour. However, the colour affects both the visibility and the penetration of the laser into the water medium. The shorter the wavelength, the better the penetration of the laser. Green lasers penetrate turbid water better than red lasers. Therefore, a green 3-line laser (520nm)[9] was selected to be integrated into the perception system. The final visual system can be seen in Figure 19.



Figure 19: Structured light system with a camera (a) and a line projector (b).

9 The main features of the laser can be found in https://www.osela.com/multi-line/

Perceptual system

Therefore, a final perception system can be built with the following sensors:

- 1x Allied Vision Mako G-125B/C GigE camera;
- 1x Gemini 720ik sonar;
- 1x Osela green 3-line laser.

The sonar is a sensor that is only used in underwater scenarios and has therefore been ready for immediate use since it was purchased. However, the camera and laser that have been selected are sensors that are suitable for use out of water. For this reason, a suitable aluminium housing was selected that would allow the system to be safely submerged and protect the sensors from some heating caused by the laser during light emission and the camera during prolonged operation.

Figure 20 shows the schematic of the final perception system.

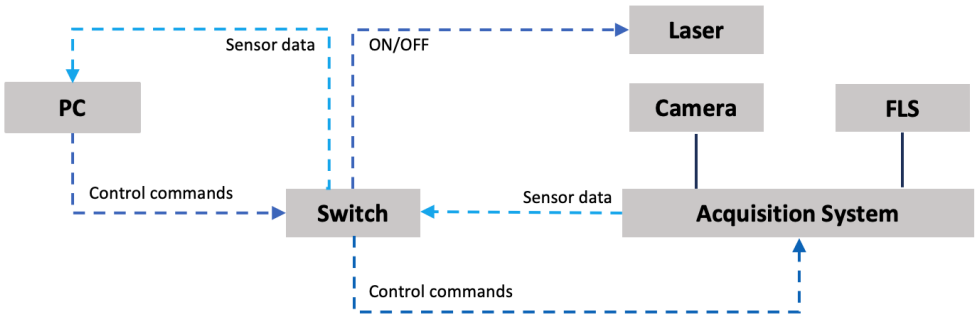


Figure 20: Scheme of the connections of the final perception system developed.

As you can see, the acquisition system block connects the camera and the FLS and supports the control and power supply of the two sensors. Via an Ethernet connection, the external computer sends commands to the acquisition system to start and stop data acquisition. A network with a switch was used for this purpose, which forwards the information between the various elements of the setup. The system also switches a laser on and off, which emits light and causes distorted lights to appear on the captured images. For this purpose, an Arduino is used, which is installed in the housing of the laser. When the perception system was developed, the idea was that the data information could be collected in an external computer and stored for later processing. Otherwise, there is an Odroid in the acquisition system that allows you to process some data online if necessary.

Customizable Remotely-operated Aquatic Bot (CRAB)

The perception system described above was integrated into the CRAB, i.e. a ROV previously developed for data collection and the inspection and monitoring of areas that are more easily accessible with a cable-driven vehicle. This allows the transport of various sensors in the underwater environment and has some interesting features, such as its dimensions and shape, which allow the robot to be used in many contexts, see Figure 21.

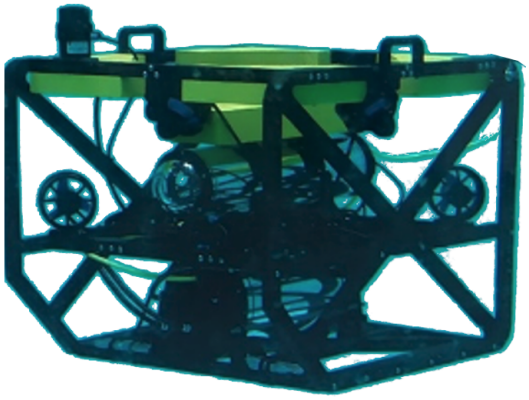


Figure 21: Customizable Remotely-operated Aquatic Bot.

The CRAB currently has a 150 metre long cable with that the vehicle can be controlled via a joystick and can be dived to a depth of 60m. It has 8 motors with that it can move with 6 degrees of freedom. As navigation sensors, it has an IMU, a pressure sensor and an altimeter. User interaction is via a control console that allows you to send commands to the vehicle and receive data. It has a battery that can power the vehicle for about 3 hours. It has some floats that allow the robot to flood when the motors are switched off and allow the vehicle to drive to the surface if something fails in the control module. This measure ensures greater energy efficiency during this process. In addition, this vehicle has two MSIS[10], that make it possible to perceive the structure of the environment, especially when there are obstacles near the vehicle, see Figure 22. It is the smallest sonar in the world and weighs only 180g in the water. For this reason, it is very widely used as it can be easily integrated into any vehicle, unlike most other sonars, which are larger and heavier. In addition, the resolution of 7.5mm is very good for the size and price of the sensor, so it is often chosen to equip these vehicles. It has a beam width of 35° in the vertical and 3° in the horizontal and can scan 360° continuously, but the scan sector is configurable up to this maximum. This sensor has a range of 0.3 to 75m and can be used up to a maximum depth of 750m. The information returned can be observed directly in its own interface or

10 The main features of this sonar can be found in https://www.tritech.co.uk/products/micron-sonar

processed by the user. The vehicle therefore has two MSIS with that it can analyse different directions: the seabed and the lateral part.



Figure 22: Mechanical Scanning Sonar used in the ROV.

It could therefore be possible to collect the desired perception data with this vehicle by integrating the perception system developed for this purpose, as can be seen in Figure 23. The laser and the camera are aligned so that the laser lights can appear on the captured images. The FLS is located on the underside of the vehicle to minimise interference when capturing acoustic data. In this way, the planned trajectories for capturing visual and acoustic data could be realised.

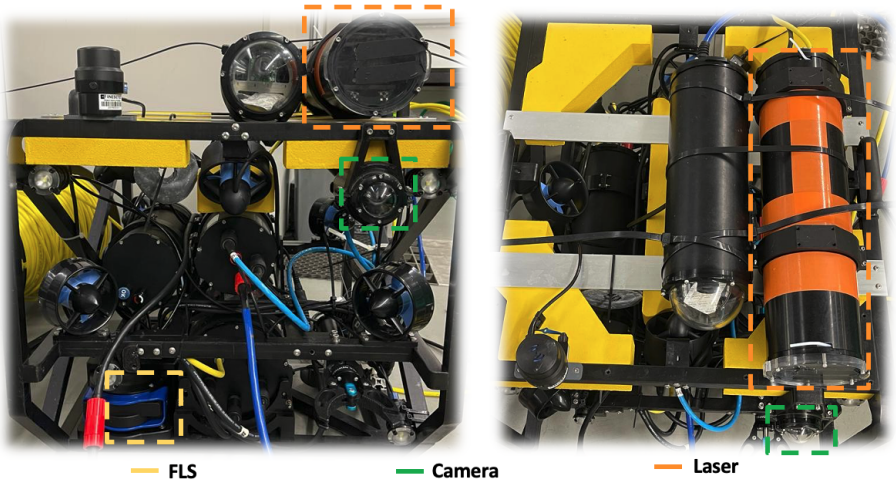


Figure 23: Perceptual system installed on the CRAB vehicle.

Underwater GPS-G2

For various tasks it is necessary to determine the position of the robot during a mission in order to compare the results later with some algorithms or even to link perceptual information with the positioning. In most cases, however, it is a challenge to determine the position of the vehicle as the methods are not always precise. Therefore, it is sometimes necessary to use external systems that track the position of

the robot. The underwater GPS-G2[11] system is a system that enables reliable and accurate acoustic positioning with a single tracking device that can be fixed in the ROV. It is a acrshortsbl system that provides robust acoustic localisation of a single tracking device. It is important to mention that it can be used in the most difficult scenarios, such as in shallow water, in tanks, etc. It has a small size that allows easy installation and is a very portable system. It consists of three basic elements: the G2 Topside, the U1 Locator and the G2 Antenna. The G2 Topside is connected to the G2 Antenna, which must be placed in the water and has a direct line of sight to the locator. The locator is attached to the vehicle after synchronisation with the topside and communicates with the antenna while the vehicle is on the water. A Wi-Fi signal sent from the G2 Topside's internal web server allows the user to check the location of the tracking device via Water Linked's web-based GUI. An integrated GPS and IMU provide an absolute position of the tracking device, so no additional third-party sensors are required. Figure 24 shows the different elements of the G2 system.



Figure 24: Underwater GPS-G2 System.

4.4 BUILT DATASETS

Inspection and monitoring tasks on the seabed are increasingly performed by autonomous robotic vehicles, and the behaviour of these platforms must be different for each object. For example, if the robot

11 The main features of this system can be found in https://waterlinked.com/underwater-gps-g2

finds a mine, this is dangerous and the robot must perform a pre-programmed action (escape or destruction). However, if the robot sees a pipeline or a chain, it can avoid the object and continue the mission. It is therefore important that the robots recognise different objects with a high degree of accuracy.

In this way, a crucial aspect in the study of the quality of underwater perception is therefore the datasets that allow the methods developed or already available to be tested according to the objectives of each application. In the underwater world, however, there are not many datasets that fulfil this need. In many cases, it is therefore necessary to create some datasets that allow some methods to be tested under real conditions. Thus, it is possible to correct some issues and problems in the development of the methods before these methods are used in real applications. In the underwater scenario, the creation of a dataset is a challenge as the data acquisition and the conditions for its realisation are not trivial. Below is a list of the developed datasets with the intended applications and the sensors used.

4.4.1 *CriticalObjectDataset*

As the existing datasets in this context are only suitable for certain objects, such as the identification of fish or coral species, or do not represent all conditions that may occur in reality, it was decided to create a dataset that could help with this assessment. Thus, for the object classification study, a dataset was developed that contains a series of images of various man-made objects that are considered critical in underwater operations. In addition, a ground truth indicating which objects are included in each image is provided to later measure the quality of the results obtained.

In this context, the following critical objects were selected for the creation of the dataset to try to represent some common objects encountered in underwater scenarios and to create an action plan that the robot can use in the future:

- Anchors and chains - These objects are usually located on the seabed and are confused with the background due to their colour, as they may be partially obscured (buried in the sand).
- Pipelines and flutuaction modules - These objects can be together or separate. The first type is grey (which makes them difficult to identify) and has some degree of fluctuation, meaning that they are not always static on the seabed. In contrast, fluctuation modules can have different colours and shapes.
- Mines- These objects have a characteristic shape and are one of the most dangerous objects in the underwater world. They can lie on the seabed or float.

- Buoys - They are usually coloured to ensure visibility for seagoing vessels, e.g. when they are used to demarcate an area in the sea. Otherwise, they can help locate underwater vehicles.
- 3D marker - This object helps with vehicle navigation, especially when tracking and docking during deployment. It can also be used to determine the 3D position of the robot. The mine and the marker were developed at a scale suitable for the shooting conditions.

The data acquisition phase was divided into two steps: clear water with water without impurities and turbid water simulated with gypsum powder, which makes the water cloudy and turbid and makes it difficult to identify the objects present in the images. The last mode was used to better simulate the real environment, as the water in the underwater environment is not clear and sediments are often stirred up by the movements of the vehicle, see Figure 25, where the effects of the sediments on the data quality are visible.

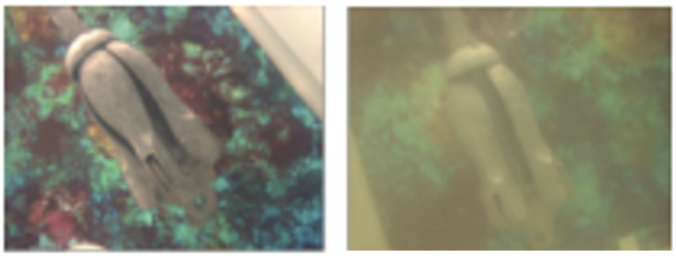


Figure 25: An example of a buoy under both conditions: crystal clear a) and cloudy b).

In both cases, the objects were recorded with a reef in the background, i.e. with a coloured background representing some reefs and real regions of the environment. In addition, all objects were also captured with a white background to facilitate training, as the white background allows the model to train the specific features of each object without any other interference. The images were taken from different perspectives of each object at 12 fps using the camera described above in the section on the perception system. They are already rectified and can therefore be used immediately. In this way, more than 800 different images per class were obtained (some classes offer about 1200 images), a number that is already useful for classification as it allows the models to be trained with diversification. They are grouped in folders by class so that these folders can be used as ground truth for classification.

Figure 26 shows some illustrated perspectives of the critical objects from the dataset under all conditions.



Figure 26: Some examples of the critical objects in different acquisition conditions.

4.4.2 WavesObjectsData

With the same focus on the observation of objects encountered during the underwater robotic mission, a new campaign of visual data was carried out in the Advanced Vehicles for Exploration of the Seas Workshop (WAVES)[12]. The main idea of this dataset is to test the object classification and segmentation approaches developed in this work. The environment shows interesting structures typical of a harbour and

12 A workshop organised by CRAS of INESCTEC, CIBIO - Azores and IS2E from 5 to 7 July 2023 in Ponta Delgada, Azores, to promote advances in marine and underwater robotics with various regional institutions and companies

some of the marine life commonly found in these places. Therefore, images of corals, rocks, fish, human divers and other robots were captured from different viewpoints and perspectives, see Figure 27.

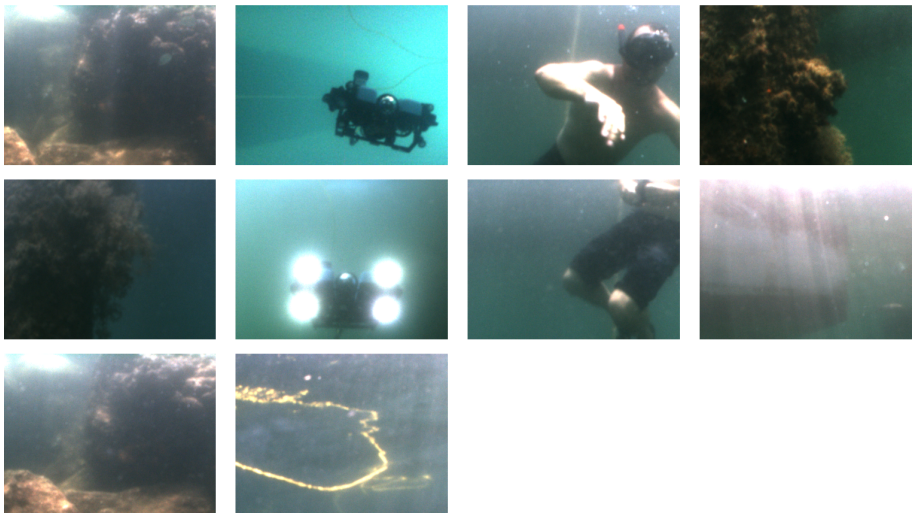


Figure 27: Some examples of the objects encountered during the WAVES campaign.

As you can see, the images show some turbidity and some difficulties caused by the sun during the acquisition, as it was carried out close to the surface. As far as the turbidity is concerned, one would have expected the water to be clearer. The fact that this is not the case is due to the work in a factory near the harbour, which affects the quality of the water. However, compared to the water in the harbour of Leixões in Porto, it is still clear. Each image consists of 640×480 pixels. In order to use these images for training the classification and segmentation models, you need to create a ground truth for these images. However, if the main goal is only to test the best training models, you can use the images as they are.

5 IMAGING ACOUSTIC MAP

This chapter presents some methods developed and their analyses to obtain representations of the underwater environment explored by a robot. This makes it possible to discuss the advantages and disadvantages of each method depending on the sensor chosen and to analyse the situations in which they can be applied.

The chapter is organised as follows: section 5.1 introduces some concepts and applications where the use of the map methods is crucial and explains how each sensor can help in this application. Section 5.2 contains a simple method to combine the information obtained from a MSIS with an external information source and demonstrates its behaviour in the map of some objects above the sensor. Section 5.3 presents a method that combines overlapping acoustic images from a FLS with the assumptions of a visual mosaic, as well as the evaluation of the obtained object representations with simulated images. Finally, the main conclusions of this chapter are presented in section 5.4.

5.1 INTRODUCTION

Mapping the underwater world is crucial for the exploration of this environment, as it is not only unknown but also unstructured. For this reason, a useful representation of the scenes is important in different areas, namely ecology, biology and geology. A good representation can be used in environmental monitoring to observe habitats and the diversity of some species in certain regions or in the monitoring and inspection of structures where a map is needed to check that everything is in order. It is also important in underwater robotic missions, namely for obstacle avoidance when the robot is investigating and exploring the environment, and for route planning, i.e. to ensure that the robot can work close to components of infrastructures. These accurate and realisable maps improve the capabilities of robotic systems and make them more effective in performing their tasks in a difficult and often dynamic environment. So, to achieve a good result in mapping, it is necessary to collect data from perceptual sensors to understand the scenes well, but this is still a challenging task. Since using visual systems in underwater mapping is a really difficult task, especially in dark environments, acoustic sensors can compensate for these limitations as they can provide information about the underwater environment even in poor or no visibility conditions. There are various

acoustic sensors that can help with this task. They need to be selected according to the application and the desired outcome as well as the information to be extracted from the final map. If the main goal is to create a 3D reconstruction of the underwater scene, an MSIS can be used, i.e. an imaging sonar, but where it is possible to extract from its raw data the distance between the sensor and the closer object. It is also possible to use an external source of information and create a 3D representation of the scene under the robot and recognise some features of the objects from it, such as the dimensions of the objects and an idea of their shape. Otherwise, if the goal is only to obtain a 2D representation, e.g. to see if there are obstacles around the robot, a mosaic can be created by combining acoustic images from a FLS along the path travelled by the robot. The pose estimation to create this representation can be derived by extracting the key visual features from the acoustic images, i.e. the same assumptions apply as for visual mosaic with optical sensors. The most important thing is that the final result is accurate and gives an idea of the structure of the scenes and in this way helps with the tasks mentioned above.

5.2 OCCUPANCY GRID MAPPING FROM 2D SONAR DATA

Three-dimensional reconstruction is an important area in underwater robotics, as there is a large amount of unknown information about this environment. However, it is a non-trivial task as it depends on the application environment and the main objective of the reconstruction as well as the information to be extracted from the final map. The success of these representations and their quality are crucial to monitor structures, check if there are obstacles near the robot and recognise some of them in the final map.

Therefore, the main goal of this preliminary evaluation is to check the behaviour of Mechanical Scanning Sonar, an imaging sonar that has a configurable rotating head (see 4.3 for more details), when mapping the scene under itself. In this way, it is possible to understand whether this unique and simple sensor is able to fulfil this task.

5.2.1 Methodology

The sensor used is capable of delivering acoustic images or two-dimensional raw data. Depending on how the sensor is attached to the vehicle, this data can be interpreted differently. However, if you want to obtain a 3D representation of the scene, you must use an additional external source of information to obtain the third dimension.

To solve this problem, three main modules are therefore usually used: one that is responsible for localisation, one that captures the perceptual information and another that is the main module and allows the accumulation of information between the different modules.

1. Localisation

One of the most important modules is the localisation source to enable correct point accumulation, as an error in the localisation estimates can lead to serious errors in the final reconstruction. In this case, the localisation source is based on a pre-developed system that enables the tracking of a visual ArUco marker. More specifically, a Blackfly S GigE camera (BFS-PGE-04S2C-CS, with a LENS-28C2-V100CS lens) was mounted on the roof above an experimental pool (4.4 × 4.6 m) to capture the images, see Figure 28.

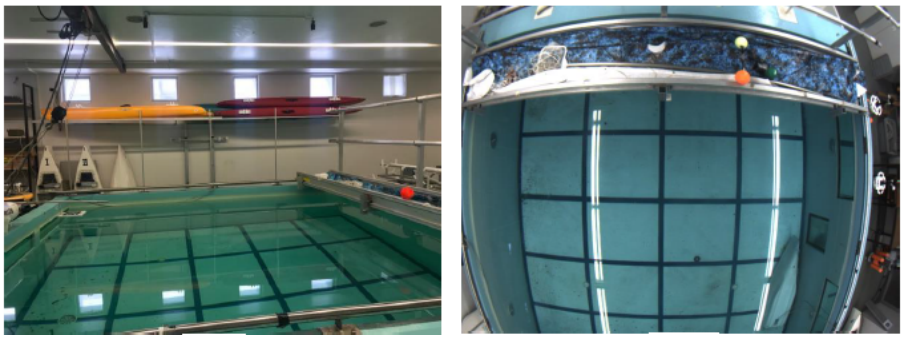


Figure 28: Setup of the camera mounted on the roof of the experimental pool for data acquisition. On the right you can see the image provided by one of the cameras.

The widest possible angle of view was used to cover almost the entire area. The camera was calibrated and the parameters obtained were taken into account during image acquisition in order to correctly extract the position data of the ArUco marker and, consequently, of the rest of the system.

Localisation based on visual markers is used for various tasks, namely in environments where other solutions are not possible, e.g. because GPS reception is difficult. Since the results obtained are very accurate, they can nowadays replace some more complex methods such as SLAM if a camera is available and one of the positions, the position of the marker or that of the camera, is known. There are many markers that are known in advance and can therefore be identified in an image. The choice fell on ArUco, which consists of black and white squares (see Figure 29 (a)), because it is simple and offers some libraries for generation and processing, making it one of the most suitable for localisation systems. In this context, it is very important that the markers are unique and invariant to orientation changes in

order to obtain robust positioning. Therefore, it is possible to obtain the position and orientation in relation to the visual system through a marker. Since the camera is about 2 metres from the water surface, a marker is sufficient to determine the position of the data acquisition system. Thus, an algorithm (based on OpenCV) was developed to identify and track the markers in order to determine the x, y position and *yaw* orientation of the marker and thus of the sonar. To do this, it was necessary to create a reference coordinate so that the returned coordinates make sense in relation to the world. In this case, as shown in Figure 29 (b), an internal corner of the test pool was used as a reference.



Figure 29: Example of a marker used (a) and the reference coordinates (b) of the determined position.

Figure 30 shows the result of the position determination in a single image, but it is important to note that this can follow the marker with 15 fps.

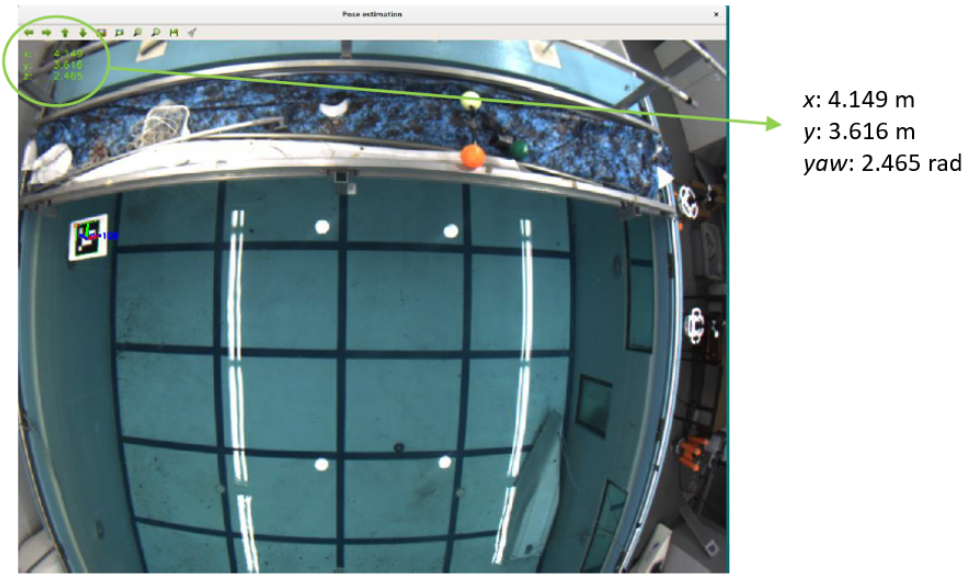


Figure 30: Result of the position achieved by marker tracking (x, y, *yaw*).

It is also known that the error of this method is between 3 and 9cm, which is quite comfortable for this type of implementation, which is about mapping medium and large objects.

2. Data acquisition

In order to test the developed method, it was necessary to create a setup that would allow the data to be recorded. The main source of information is therefore a MSIS from Tritech, which is known for its configurable Field-of-View of up to 360° (where a larger FOV means slower data acquisition). In this case, the FOV has been set to 30°, i.e. 15° for each side. This sensor has a low horizontal beam of (3°) and a high vertical beam of (35°). This means that the reflected echo may have originated anywhere along the corresponding elevation arc, which is considered a disadvantage of this sensor.

In Figure 31 you can see the final setup, i.e. the sonar is attached to a platform in a horizontal configuration to be able to map the area below the sensor and the visual marker is attached to the surface of the platform to obtain the localisation.

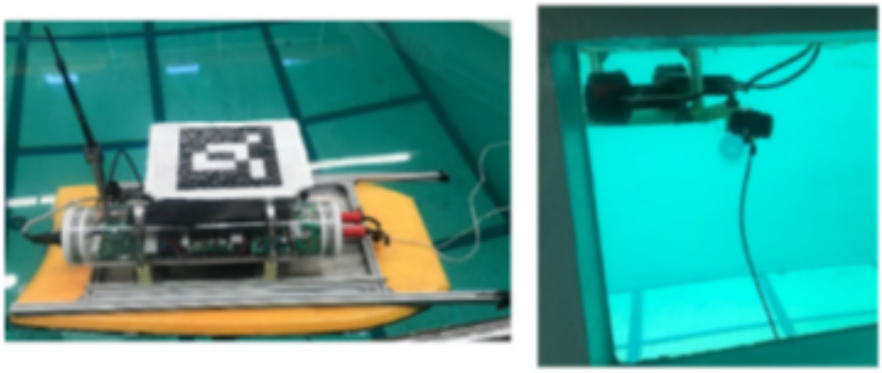


Figure 31: Setup configuration for data acquisition with an ArUco marker and Tritech MSIS.

3. Pointcloud accumulation method

Nowadays, it is also important to think about memory requirements, as this is very expensive for this type of application. This consideration is important, for example, when the data is collected and sent to a centralised system that creates the map and needs to send this information to the main vehicle so that it can make some decisions (without sacrificing storage space).

There are many approaches to solving the mapping problem with acoustic sensors, but usually occupancy grids are used. These structures represent a probability for the occupancy of a space within an environment. In this case, they are useful because the Mechanical Scanning Sonar have an inherent uncertainty due to the sensor model

and therefore they help predict whether the cell is occupied or not. In this way, the sensor readings can provide insight into whether or not there is an obstacle in a scenario. For this purpose, the probability resulting from the fusion of the cumulative measured values of a specific position is used. This probability is updated together with the data acquisition and mapping process. In this way, some tools are already implemented and available in the literature that use the occupancy map approach in a 2D context or even for a 3D map. For example, it is possible to use the MATLAB Navigation Toolbox or the Octomap library [84]. Both implementations are useful tools as they allow the creation of a map by assigning to each cell - independently - a probability indicating whether this cell is occupied or free [85]. In this way, analysing the final map is simple and more intuitive. Although the proposed method has already been implemented with both tools, the final results are demonstrated with the MATLAB toolbox. To solve this mapping problem, a point accumulation method was implemented, and so the technique can be divided into four phases, see Figure 32.

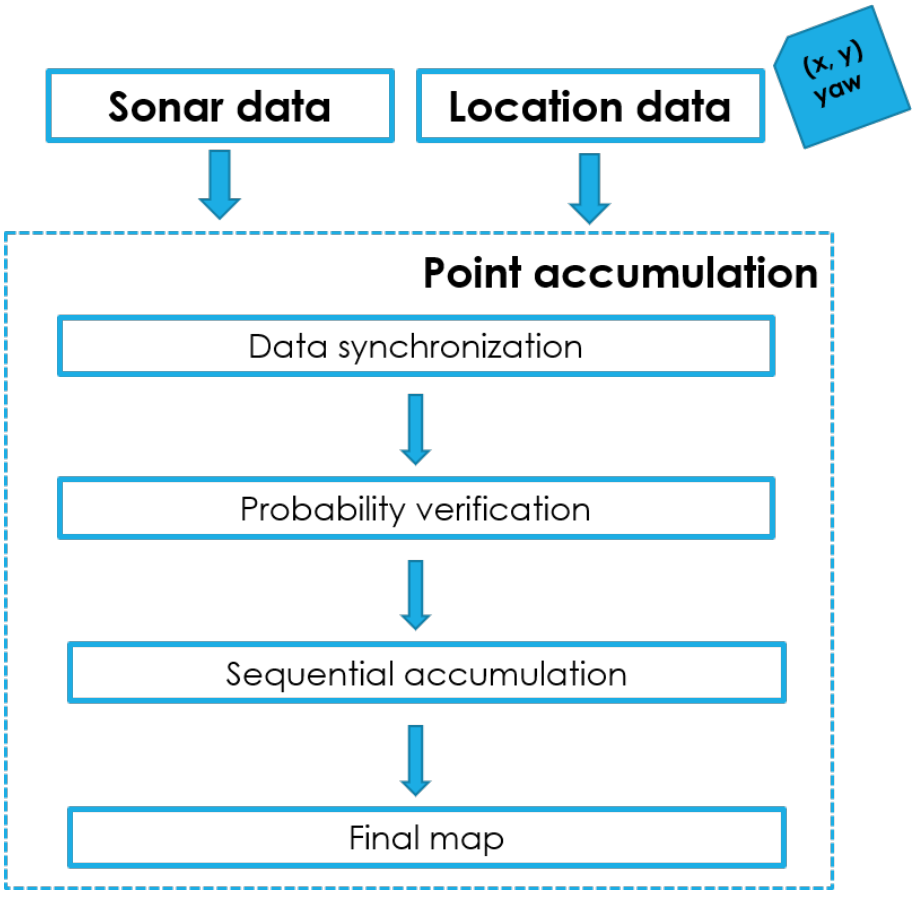


Figure 32: Overview of the developed point accumulation method.

After acquiring the sonar and localisation data, it turns out that the two sensory information are not synchronised, as the frequency of the two is very different (the sonar has an acquisition rate of 24 fps, while the localisation data has 15 fps). When incorporating the information

into the method, it was therefore necessary to compare the timestamps of the two pieces of information, which means that each sonar sample is linked to the most appropriate position value. The sonar data must then be filtered, i.e. the artefacts caused by the sonar housing are deleted. Then the point accumulation is started, selecting the strongest bin of each tick. Once the 2D position (x and z) with respect to the sensor is known, you can assume that the 35° arc is fully occupied, as the reflected echo can come from anywhere. Nowadays, this approach is a common solution to this problem in modern methods. However, this assumption implies a maximum error higher than the expected one (over 55cm), which could lead to some errors in the prediction of the dimensions of the objects and their position. For this reason, it was decided in this work to attribute a lower probability of occupancy to the extremities of the arch than to the central part. This method therefore eliminates about 40% of the entire arc, as the probability that the reflected eco is located in the central part is higher. Therefore, the correction was made so that only the central 20° of the arc is retained. This adjustment improves the results as the maximum and constant error of the object dimensions decreases to 33cm, which is a more practical value, and consequently the error in the pose estimation of the objects also decreases to 15cm. Next, two different tests are carried out to ensure that the map information is the most relevant:

- **intensity** - the main objective is to check whether the intensity of the selected strongest bin is higher than a user-defined threshold. Therefore, the bins with lower intensity values (less than 125 in a range from 0 to 250) are discarded as their information is not considered relevant. This procedure is always applied to all returned echoes.

- **scan angle** - this test makes it possible to assign different probability values to each point depending on the corresponding scan angle. Therefore, the points corresponding to the scan angle of the extremities of the FOV have a lower probability of occupancy than the points corresponding to the central angles. According to a previous study, the selected boundaries are chosen to cover more than 75% of the FOV, see Figure 33. It is important to note that this filtering can be enabled or disabled depending on the object size, i.e. when reconstructing smaller objects, this filtering must be disabled.

These two checks reduce the final number of points in the reconstruction without affecting the quality of the map.



Figure 33: Probability assigned to each sample according to the scan angle when the sonar is configured with an angle of -15° to 15°.

Table 4 shows the results of an experiment on the hull of a ship in terms of the number of points based on the two filtering methods. Firstly, the intensity test reduces the final number of points a little. If you use the intensity and scan angle check at the same time, you can see that the final reduction is more than 35% of the total points. In this way, it is not necessary to sacrifice memory space because the quality of the map is similar.

Table 4: Comparative analysis of the number of points without and with the filter techniques.

Original data	Intensity filter	Intensity & Scan Angle filters
4494	4095	2744

The data can be accumulated after all test procedures. To do this, it is only necessary to bring the sonar data into the correct position provided by the localisation module and thus into the world reference. In this case, the coordinates of a test pool are used as a world reference. An octree, which is widely used in various applications, was used to accumulate the points, as it offers an efficient way of organising and querying spatial data hierarchically. The chosen storage structure was created with a resolution of 100 cells per metre, which means that each cell is 1×1 cm. It was also determined that only the points with an occupancy probability above the standard toolbox threshold - 0.65 - should be included in the final map. Therefore, the final map is updated cyclically as new information is received to identify areas with obstacles near the sensor.

5.2.2 Experimental Results

Developing a method for a real-world application is a really difficult task as it needs to be tested in different ways, under different conditions and in different scenarios to ensure that the method is able to achieve the main objectives proposed. Therefore, an evaluation is performed in terms of building materials that vary the required sonar

gain and some scenarios are used to evaluate the proposed method in terms of tested dimensions and observable shape. Finally, a discretisation approach is also tested, which allows to reduce the memory requirements without losing the most important information. The evaluation is done through some experiments on a test pool with some objects with different properties.

When a robot navigates in an underwater environment with the main goal of inspection, it encounters different objects with different materials. In order to obtain a first sensitisation of the returned echoes, the first experiments were performed with the aim of comparing different values of gain. The gain was varied between 30% and 60% for 5 different materials - wood, PVC, metal, concrete and plastic - see Figure 34.

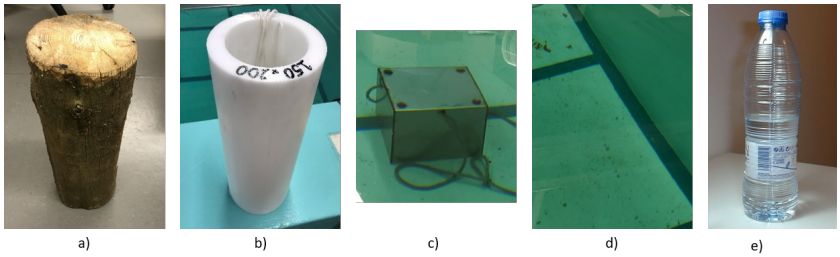


Figure 34: The 5 types of material used to evaluate the sensor: wood (a), PVC (b), metal (c), concrete (d) and plastic (e).

Figure 35 illustrates the difference between the two objects: a metal box and a plastic bottle in the MSIS software. The plastic bottle is made of a material that does not provide useful echo values. This fact can be explained by the small thickness of the water bottle. The difference in density between this material (when the bottle is filled with water) and the water is small. Therefore, the echoes are absorbed, see Figure 35 b). In contrast, the metal echoes return with higher values from a gain value of 20% (see Figure 35 a)).

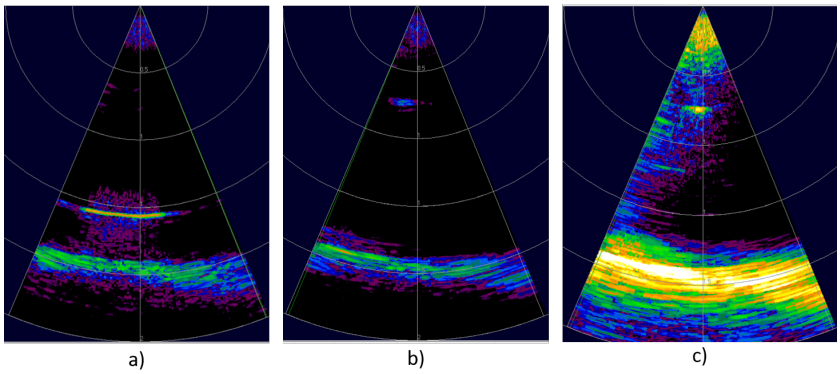


Figure 35: A metal box a) and a plastic bottle b) were recognised with a gain of 20% in the software of MSIS. A plastic bottle with 70% gain is also shown c).

The bottle is closer to the sensor, but the values returned at 20% of the gain parameter do not provide useful values. In an extreme test with 70% of the gain, the bottle is more visible, but there is a lot of noise and the scene is generally saturated (see Figure 35 c)). However, this type of plastic is not easy to find in the sensor's FOV. This is because when the bottle is filled with water, it floats, and when it is filled with sand or sediment, it is near the bottom (i.e., it is similar to a deposited sediment). In addition, to analyse the sonar data and select the best value for the gain in general, different values are tested with different trajectories over the metal box, the PVC cylinder, the concrete and the wood trunk. Table 5 summarises the results obtained during a mission with the MSIS driver software in terms of returning echoes. Most materials are clearly visible with a gain of 40%. However, in order to choose a single value for all materials so that their visibility is not affected by the filtering techniques of the proposed method, a value of 50% gain has proven to be the most suitable for these materials.

Table 5: Results in relation to the returned echoes with variation of the gain parameter.

Material	**30%**	**40%**	**50%**	**60%**
Metal	[110,135]	[125,135]	[150,165]	[160,175]
PVC	[95,110]	[110,125]	[130,140]	[140,155]
Concrete	[90,130]	[125,140]	[130,150]	[140,155]
Wood	[105,115]	[125,145]	[135,160]	[165,180]

In order to test the developed method, it was necessary to collect several sets of localisation and sonar data. Each set consists of a file with the sonar data and the corresponding timestamp and another file with the localisation data and also the corresponding timestamp. To demonstrate and comment on the main results, different objects were selected to prove the robustness of the method as well as some of its limitations. A metal box, a trunk of wood and a boat hull (where the sonar behaviour resembles that of PVC in terms of reflectivity) were selected to test whether the method can correctly determine the dimensions of the objects and whether it is possible to recognise the shape of more complex objects and with non-straight shapes, see Figure 36. In this way, the scenarios used with these objects and some characteristics of each are visible in Table 6.

The first experimental test (scenario A) was carried out with a metal box, as it makes it possible to check the dimensions of straight objects. As far as positioning is concerned, the consecutive samples are correct and well related to the experimental tank. The *yaw* alignment had to be filtered as the consecutive values show some inconsistencies, as you can see in Figure 37 a). The result of the application of this filtering is shown in Figure 37 b).

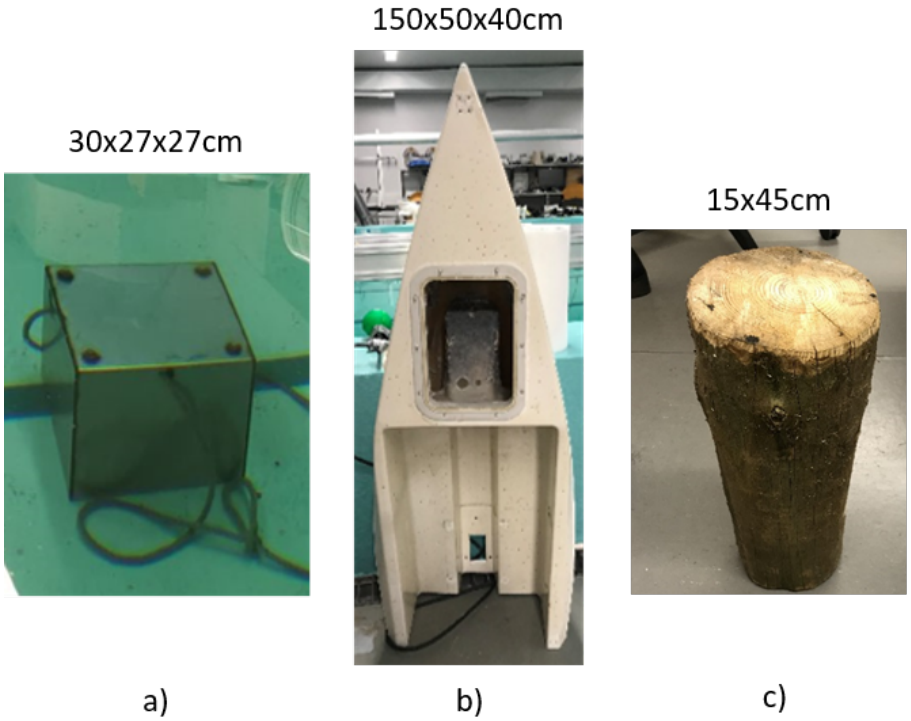


Figure 36: Objects used to test the developed method: a box (a), a boat hull (b) and a trunk (c).

Table 6: The scenarios used to test the developed implementation.

Objects	Scenario	Trajectory	Velocity [m/s]
Metal box	A	y direction	0.08
	B	y direction	0.13
Trunk	C	y direction	0.03
Hull of boat	D	x direction	0.04
	E	x and y direction	0.05

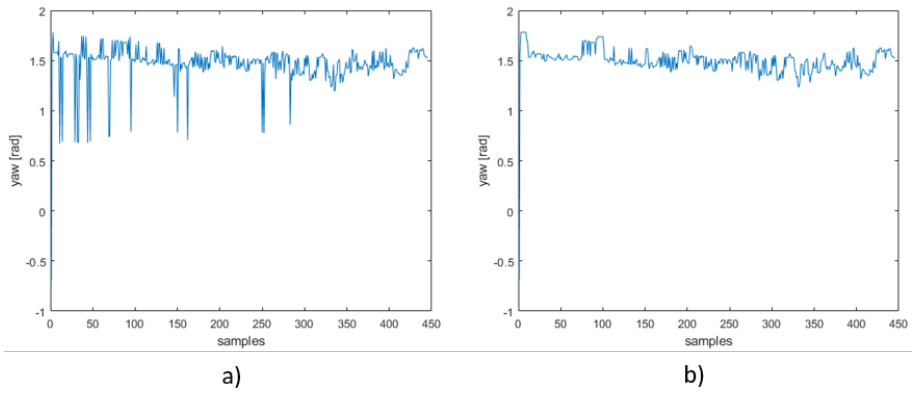


Figure 37: The original *yaw* angle a) was filtered to ensure consistency along the trajectory b).

Regarding the accumulation points, it is important to point out that the scan angle filter has been disabled in this case because the metal box is small and it is crucial not to delete important data. The dimension in the direction of movement (y) determined directly with the method is 59cm. As explained earlier, due to the assumption that the vertical arc of the sensor occupies more than one point, there will always be an error in the direction of the predominant motion. In this implementation, it is assumed that only the central part is occupied, as the probability of this event is higher. Therefore, the detected size of the object has a constant deviation of 33cm in the y-direction. In this direction, the proposed method recognises a box with 26cm. However, in the other directions (x and z direction) this assumption is not necessary. As you can see in Table 7, the obtained dimensions are well estimated for all dimensions of the metal box, with a maximum error of 4cm.

Table 7: Dimensions obtained of the metal box in scenario A.

Scenario A	**x dimension [cm]**	**y dimension [cm]**	**z dimension [cm]**
Method	23	26	25
Original	27	30	27

The second test with the metal box (scenario B) shows similar results to the previous one. It is important to point out that the speed of data acquisition is crucial and alters the results obtained, since this set is acquired at a higher speed and for this reason only two scans represent the surface of the box. In this way, although the dimensions are correct, it is difficult to recognise the shape of the object, e.g. when the speed is high.

The reconstruction of the wood trunk performed with scenario C shows that the object, see Figure 38, is 45cm above the ground. This measurement is correct because the trunk was 15cm thick, but this object fluctuates over 30cm above the bottom of the test pool. The dimensions estimated by the method in terms of x and y are correct, as you can see in Table 8. It is important to note that the scan angle filter was also disabled in this test.

Table 8: Determined dimensions of the trunk in scenario C.

Scenario C	**x dimension [cm]**	**y dimension [cm]**
Method	20	42
Original	15	45

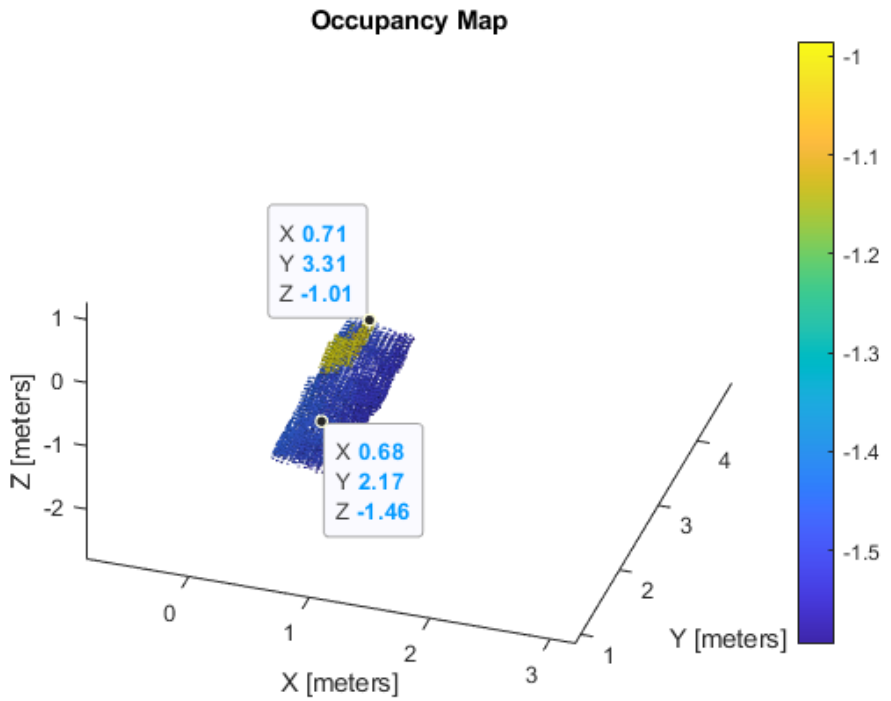


Figure 38: Final map of the surface of the trunk and verification of the dimensions in relation to the height.

Scenario D was executed over the surface of the boat hull and shows that the complex shapes are visible in the final map. To create the map of this object, the scan angle filtering is activated because the hull of the boat is larger than the others and therefore it is possible to maintain the shape of the boat on the map without affecting the storage cost by increasing the number of points. As can be seen in Figure 39 (top), the shape of the hull is well generated and it can be seen that there is a funnel-shaped part and a larger part. As for the dimensions, the length of the mapped boat is well estimated (1m, taking into account the maximum error of 33cm), see Figure 39 (bottom). It should therefore be noted that the boat is not fully mapped in terms of length and width (only 40cm), as the survey carried out deviated from the centre of the boat.

The last experiment (scenario E) was performed with a change of direction with respect to the x and y movement and the filtering of the scan angle is also activated. As you can see in Figure 40 (top), the funnel shape of the boat is visible in the final reconstruction, as well as the majority of the boat. The blue area is the bottom of the test pool, which is about 1.5m away from the sonar. This is a correct indication as the pool has a height of 1.70m and the sonar is over 20cm from the surface. As for the y coordinate, Figure 40 (bottom) shows that the estimated coordinates are correct, as the map has about 40cm in the largest part and 23cm in the funnel part.

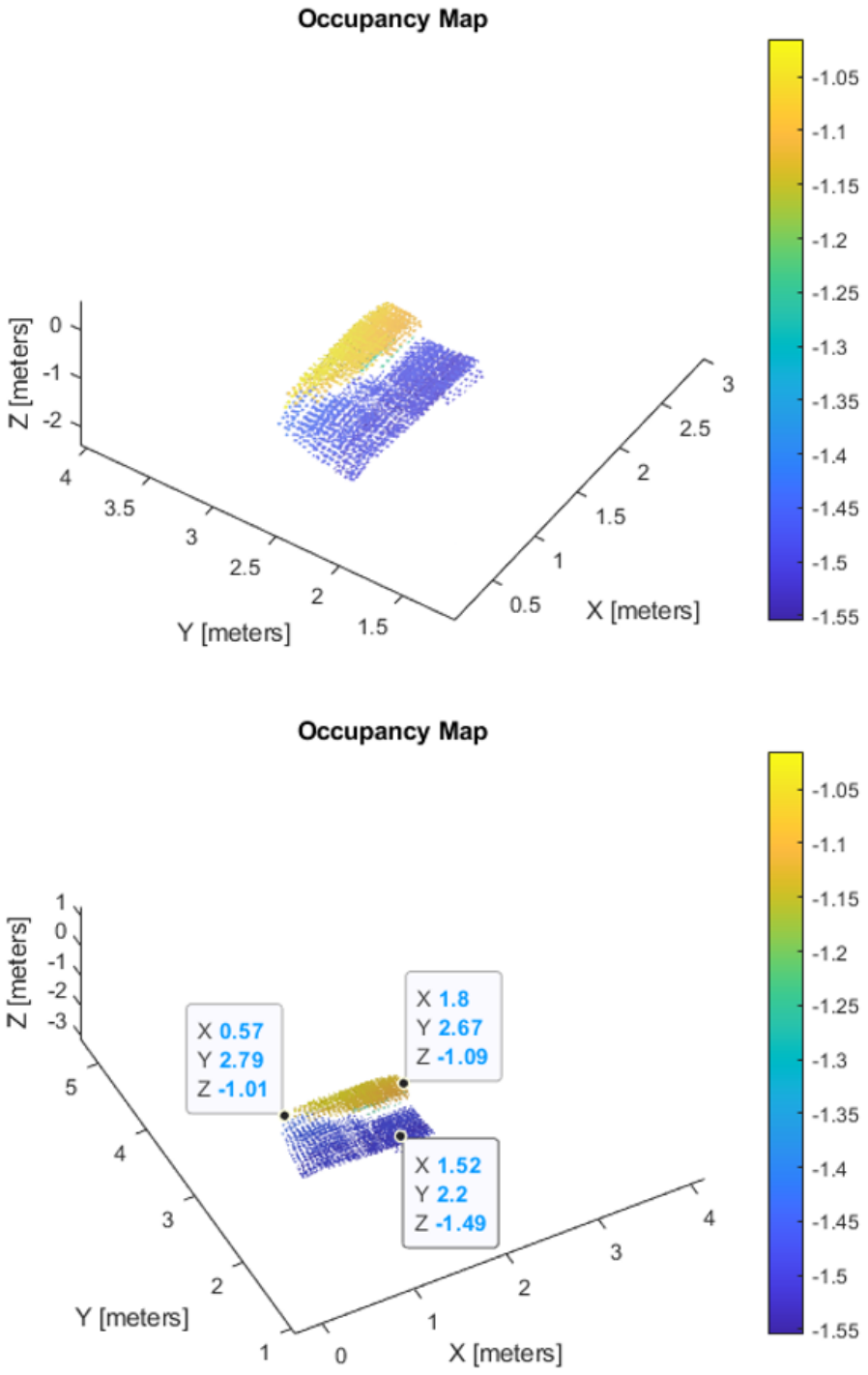


Figure 39: The final map shows the shape of the hull (top) and the verification of the dimensions in x and z coordinates (bottom).

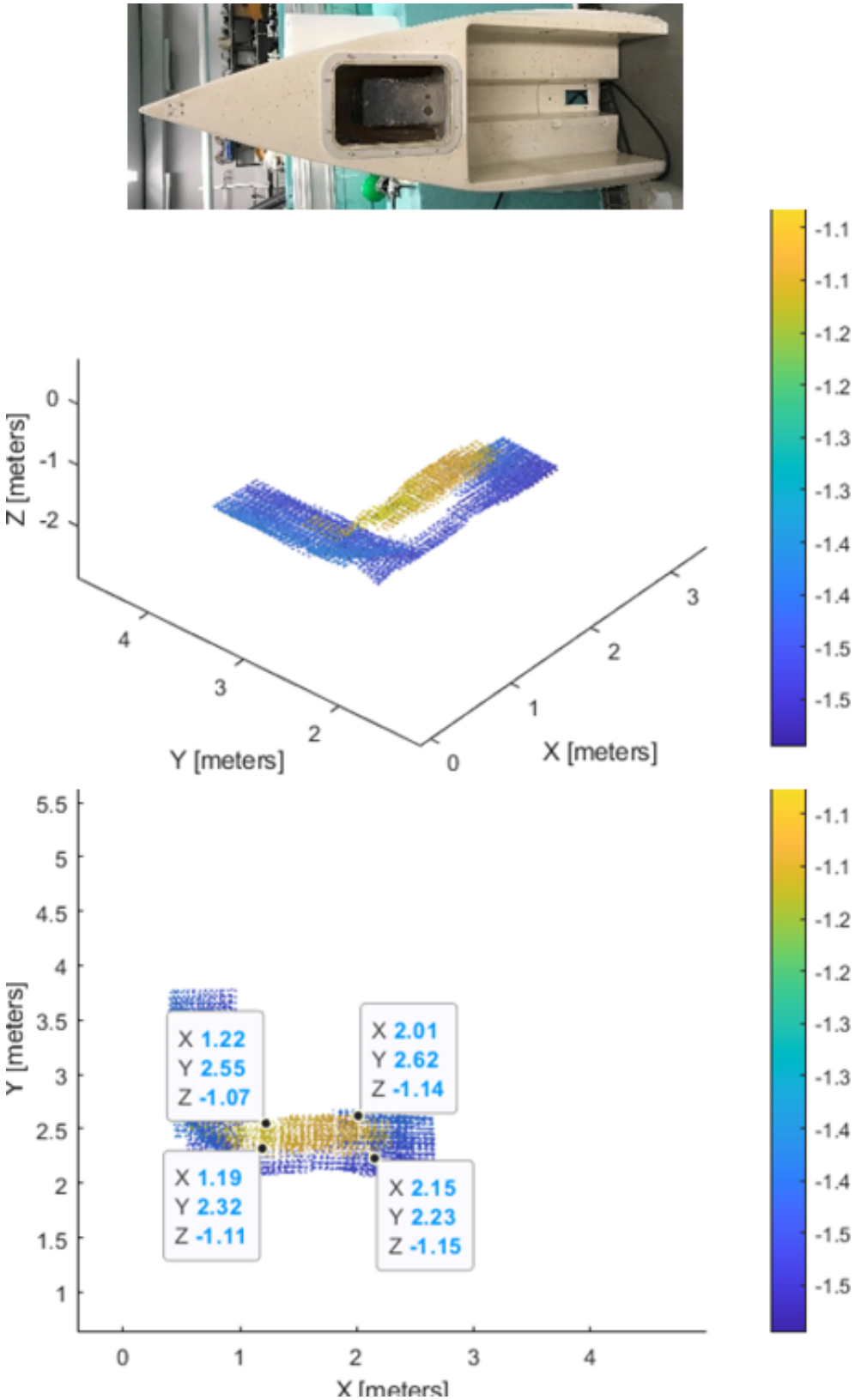


Figure 40: Final reconstruction of scenario E (top) and verification of the dimension in the y coordinate - width of the boat (bottom).

There is a small error in the largest part (7cm), which could be explained by a deviation in the localisation method or even in the filtering technique. As for the length (x coordinate), it is correct because we see that it is about 1.5m (after subtracting the maximum possible error of the method), which corresponds to the true measurement.

Sometimes, in real applications, it may be necessary to send only the points of the fixed grid to the control module. To test this function, it was therefore necessary to binarise the map obtained with respect to the coordinates x and y with different step sizes (cm). The original values, i.e. the sonar measurements, are retained for the heights. Table 9 shows the number of points for the tested step sizes for scenario E, which has an original number of 10318 points.

Table 9: The number of points after the discretisation process for scenario E.

Step size [cm]	Number of points
1	7664
3	2789
5	1125
10	465

From this table it can be seen that the total number of points is greatly reduced, which can be an advantage for real-world applications as the final map retains the shape of the objects. The step size depends on the application of the final reconstruction and must be chosen by the user. Figure 41 shows the resulting map with a step size of 5cm.

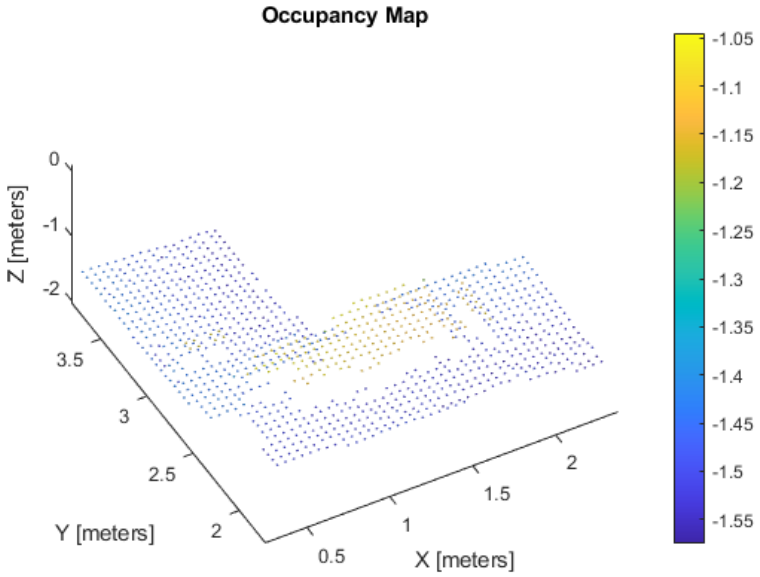


Figure 41: Discretisation process applied to scenario E with a step size of 5cm.

The selected step size reduces the vehicle's memory requirements by around 90%. In addition, good results are visible in terms of the displayed shape without loss of information. Sometimes a small error (in the order of cm) can be seen when checking the dimensions, as this method is an approximation to the real map in terms of the coordinates x and y.

5.2.3 Discussion

Thus, a simple method was developed to obtain 3D reconstructions of the scenes under the sensor by using external source localisation and octrees. This made it possible to evaluate the performance of a simple sonar, such as MSIS, in mapping task and to recognise that objects are in the vicinity of the robot. One of the most important features of the method is that it performs two different filtering techniques: one according to the returned intensity and another according to the scan angle (which can be switched off). In addition, unlike the usual techniques, only the central part of the 35° arc of the sensor is considered occupied (about 20°). These techniques make it possible to reduce memory consumption when creating the map without losing important information. From this it can be concluded that:

- About the experiments with different materials: Most of the tested materials are clearly visible with 40% gain. However, in order to choose a single value for all materials so that their visibility is not affected by the filtering techniques of the proposed method, the authors considered a value of 50% gain to be the most appropriate.
- If only the central part of the arc is occupied, the constant error in the measurement dimension of the direction of movement is reduced to 33cm (instead of 55cm).
- The method makes it possible to map smaller objects with straight lines without having to resort to the scan angle filter. The estimated dimensions in the final map have a maximum error of 5cm, which is a good result for real-world applications.
- The developed techniques show that it is possible to observe objects with more complex shapes and create maps with trajectories with directional changes. For larger objects, it is possible to manually activate the filtering of the scan angle and thus reduce the number of total points. It is also important to point out that a more correct estimation of the positioning contributes to a better mapping of a scene.

- The z dimensions of the objects are correct and the MSIS shows that it is suitable for this type of survey, with the main objective being to find the obstacles near the sensor.
- The surveying speed proves to be important, as some information is lost with faster trajectories. On the other hand, the configuration of the scan angle can also affect the quality of the final map, because the larger the FOV, the slower the rotation speed of the sensor head. A balance must therefore be found between the FOV and the speed of the vehicle.
- Finally, performing a discretization process with different step sizes shows that it is possible to achieve a simpler representation of the scene that is able to collect some information about the objects and with less storage cost. This process is crucial, for example, in cases where the data is collected and sent to a centralised system that creates the map and needs to send this information to the main vehicle so that it can make some decisions (without sacrificing storage space).

5.3 MOSAICKING APPROACH WITH ACOUSTIC IMAGES

Mosaicing is the process of combining different images with the same region to create a larger image without seams. It is often used for specialised tasks, such as aerial photography or mapping specific underwater regions. Therefore, it was possible to combine several images as they have an overlapping area, see Figure 42.

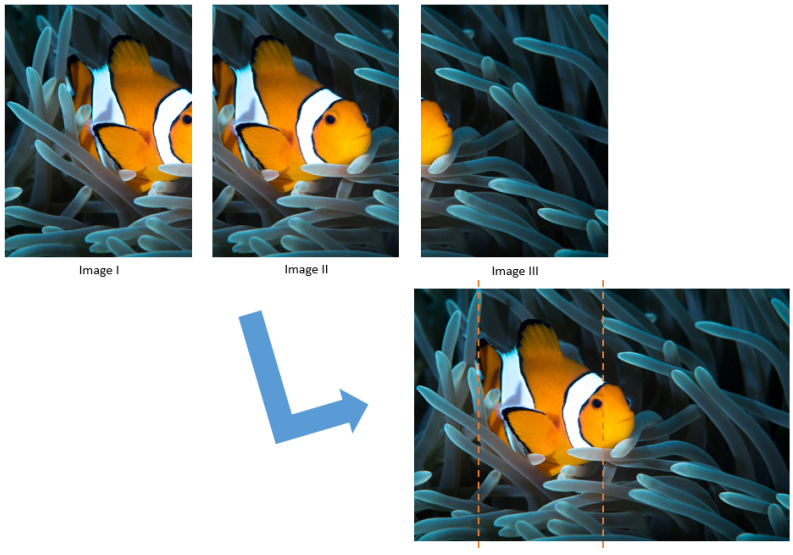


Figure 42: Illustrative example of a mosaic after the processing of 3 images.

The main objective is to obtain images that are as coherent and visually consistent as possible, corresponding to the movement of the camera, but with a higher degree of coverage than a single image. This technique is mainly used for visual images, as there is already a work on it [30]. However, the idea is to test the same assumptions with FLS images, which contain a lot of appearance information but have some problems that are not present in visual images.

5.3.1 Methodology

To obtain a mosaic image, you must compare two images, align the second image with the first and insert the second image into the first so that the overlapping area appears only once. It is also important to point out that the success of this technique depends on the success of the individual steps. However, to apply this technique to acoustic images, you need to make some assumptions, such as:

- The acoustic image must be cropped to a ROI (rectangular area) as described in section 4.2.
- The consecutive images must have an appropriate overlap area between the individual images so that they can be merged (approx. 40%);
- The sensor has no oscillating motion, i.e. the reference in the z axis does not change when the images are captured.

This requires many steps, which are calculated iteratively. The individual steps and the configurations used are explained below. There are different ways to divide the main steps and different names for each one, but in this work it has been divided into 3 mandatory steps: Feature extraction, registration, stitching and another that is optional: memory management, see Figure 43.

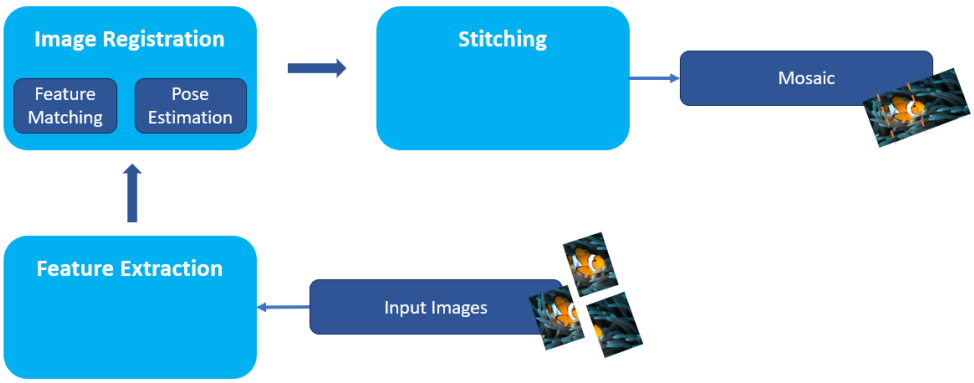


Figure 43: Pipeline of the implemented mosaic method for the application to FLS images.

Feature Extraction

Normally, the first step in assembling a mosaic is to perform preprocessing to improve and standardise the quality of the images. However, in this case, this is not applied because in acoustic images, the appearance of the objects does not change when some parameters of the environment change, such as the inclination of the sun or the turbidity level of the water. Therefore, feature extraction in both images to be merged is the first task and one of the most important steps, as the success and quality of the output mosaic is directly related to the success of this task. In this step, the most important features in both images are detected. These features are the keypoints with more information and more distinctiveness that characterise the elements of the image. It is therefore crucial that they are as reliable as possible to clearly identify the elements of an image, since the main goal is to look for the same points in the second image.

According to the literature and knowledge of this process in visual images, there are several feature extractors and given the need for viewpoint invariance combined with the high time cost of traditional feature detectors, binary methods are increasingly used for similarity detection. They offer several advantages, such as lower memory and computation time requirements, as their representation is compact and they are used for real-time applications. Therefore, it is important to evaluate their behaviour in this type of application. Thus, binary feature extractors generate binary descriptors that represent local image features, but there are several options such as Binary Robust Independent Elementary Features (BRIEF), Binary Robust Invariant Scalable Keypoints (BRISK), Oriented FAST and Rotated BRIEF (ORB). BRIEF generates binary descriptors by comparing the intensity differences of pixel pairs. BRISK is designed to be more robust to scaling and rotation and ORB combines the Features from Accelerated Segment Test (FAST) keypoint detector (which mainly detects the corners present in the image) with the binary descriptor BRIEF. Therefore, only ORB and BRISK are tested and compared, as the number of keypoints returned can vary greatly when one feature extractor is compared with another.

Registration

The next step is to compare the features detected in each image and determine which keypoints are present in both images, because it is on the basis of this selection that the rest of the algorithm works. For this we can use different matchers, such as the brute force matcher, which calculates a distance between the keypoints and determines the best matches based on certain criteria, e.g. the smallest distance or whether a distance threshold has been reached. Another example is the Fast Library for Approximate Nearest Neighbors (FLANN),

which approximates each point to its nearest neighbour and is very efficient. For binary features, you can use the brute force matcher, which calculates the distance between each descriptor in the first set and each descriptor in the second set. However, since the descriptors are binary, it is not possible to use all types of distances. In this case, the Hamming distance was used, which measures the number of different bits between two binary strings. The matcher is then applied to the descriptors and the matches must be sorted based on their distances. Often the number of keypoints is higher, but the features are not very relevant, so the matches returned are not precise. For this reason, the selection of the best matches is also crucial and so a test ratio based procedure was applied.

Therefore, the first aim of this work is to evaluate which of the two binary detectors (ORB and BRISK) performs best in recognising features in acoustic images, based on the number of keypoints but also the good matches. For this study, 3 different regions of the *Port_Mosaic* were selected, referring to sand, metal grid and stones, see Figure 44.

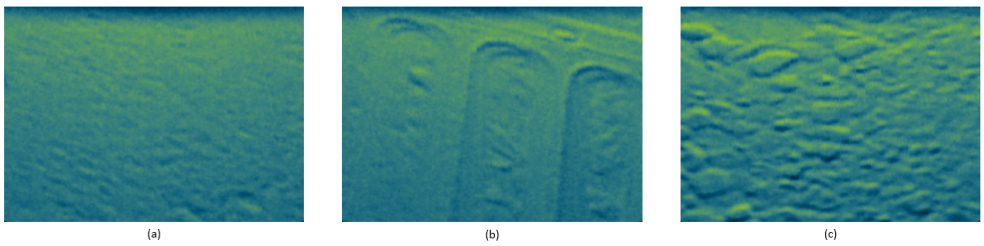


Figure 44: The different regions of the *Port_Mosaic* Dataset: sand (a), metal grid (b) and stones (c), presenting different textures.

In each region, 10 consecutive images were selected and the two feature extractors were applied with the same configurations in terms of the required feature extraction to allow comparison. The keypoints are then matched and the matches are further filtered with the same constraints. Table 10 shows the main differences between the results of the two extractors for the three different regions.

Table 10: Average number of keypoints and matches in 10 consecutive images in three different regions: Sand, Grid and Stones.

	Feature Extractor	Number of keypoints	Number of matches
Sand	ORB	3063	24
	BRISK	2642	14
Grid	ORB	3015	42
	BRISK	2638	19
Stones	ORB	3156	363
	BRISK	2926	278

Based on the results obtained, it is easy to see that the number of keypoints is always lower when using BRISK, even with a high number (in all cases more than 2500 features per image). However, as already mentioned, a high number of keypoints does not always mean a good result in the matching process, i.e. many of these points are discarded during the matching process as their robustness is low. The analysis of the results shows that the ORB feature extractor provides a higher number of matches with the quality in the ratio test. Visually, you can see in Figure 45 that the returned matches between the same images for the ORB (Figure 45 (a)) and the BRISK (Figure 45 (b)).

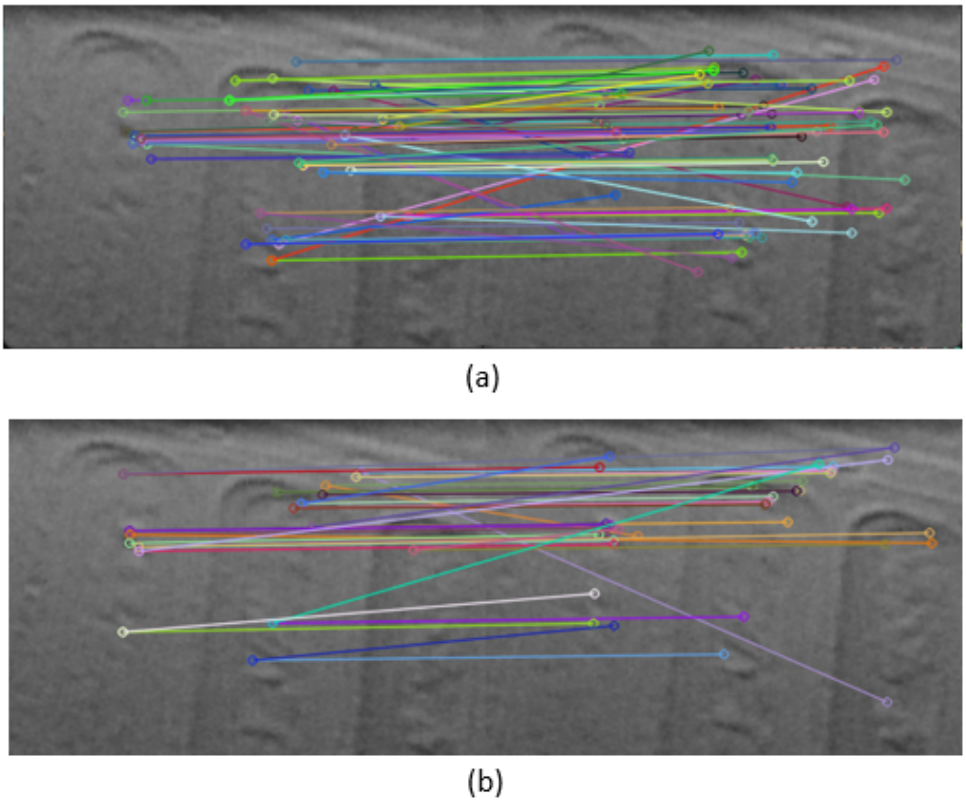


Figure 45: The differences between the matches obtained with ORB extractor (a) and BRISK extractor (b).

Thus, it can be seen that the ORB has more correspondences and higher quality as it has more straight lines, which means that the distance between the points in the two images is smaller, resulting in greater robustness and indicating that the points are considered to be the same with greater certainty. Therefore, this feature extractor was selected for the next experiments.

The second step of the registration is to estimate the transformation between the two images based on the previously extracted and selected best features. There are common transformation models such as affine transformations, projective transformations (such as homographies) and non-rigid transformations. In this case, homography has been

used to map points from one plane (first image) to another (second image), usually between 2 different images with different viewpoints but with an overlapping area. Therefore, it is possible to obtain the matrix H, which is a 3×3 matrix that correlates the coordinates $(x, y, 1)$ of a point in one image with the coordinates $(x', y', 1)$ of a corresponding point in another image with the following equation:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix},$$

where the elements h_{ij} are the parameters of the homography matrix.

In this case, the Random Sample Consensus (RANSAC) method was used to calculate the homography, as it has proven to be a robust estimation algorithm, even if there are some outliers.

Stitching

In order to merge two images, the perspective of the second image must be changed in relation to the first image. To do this, the result of the homography estimation from the previous step is used. This process, called warped perspective, is a series of mathematical transformations that allow the second image to be repositioned in order to obtain a new output image with larger dimensions. This new size means, for example, that the images are shifted to the right. In this case, the new size for the width must therefore take into account the size of the first image, the part representing the new information of the second image and the change in perspective.

When using outdoor or underwater images for this task, an additional step is usually used that allows the final mosaic image to be seamless, i.e. the seams of the combination of different images are not perceived. This step, known as blending, ensures a smooth transition between the images. In many cases, these seams indicate that the visual conditions have changed during the acquisition, e.g. due to light or shadow. Since the main goal of this work is to test the visual approaches in acoustic images, this step is not considered since the appearance of the objects in these images does not change over time. Furthermore, this approach, if applied in a conventional way, could lead to errors as the colours of the images indicate the distance between the sensor and the object, i.e. the height of the object in question. However, when testing the method, it was found that with large perspective changes between the images to be combined, some black areas appear in the centre of the final mosaic, as the overlapping areas are not blended. So when we merge the two images, the coloured areas remain coloured, but the black areas created by the warping perspective appear, see Figure 46.

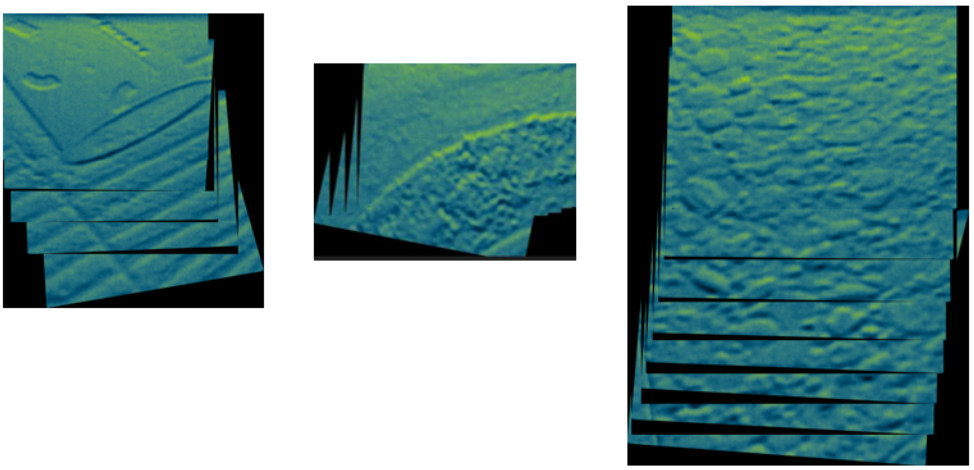


Figure 46: Example of some finished mosaics with areas with black pixels that lead to an unsightly result.

In order to avoid these incoherences, which complicate the interpretation of the final mosaic, a different approach was introduced. This consists of carrying out a cropping step after each combination up to the last image. This cropping preserves only the most important part of the mosaic without losing the idea of the movement of the vehicle. To better understand the idea, see Figure 47. The result is that the black areas do not appear and the final mosaic is more visually appealing.

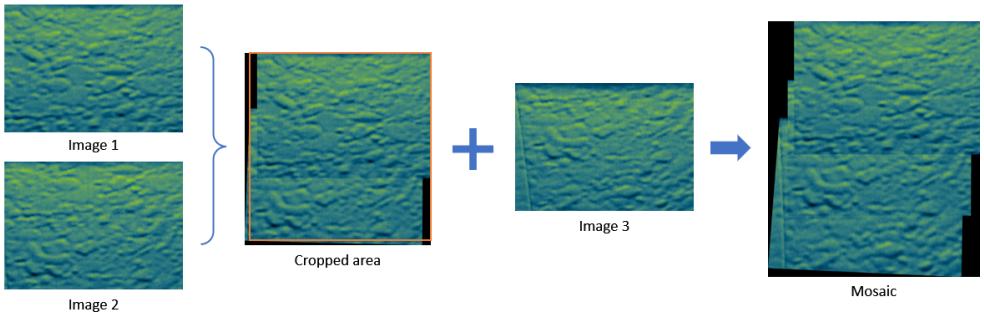


Figure 47: Illustrative representation showing the introduction of the cropping method when combining three images.

5.3.2 Experimental Results

The development of a method for a real context is a difficult task and therefore it is necessary to test it well under different conditions in order to verify the quality of the result. Therefore, an evaluation of the performance of the proposed technique is carried out. These experiments are first performed with the acoustic FLS images captured in a Stonefish simulator. These images are cropped to obtain a rectangular ROI and to be able to apply the previously described steps. A complex

trajectory is created and some sections of it are used to test the method. More detailed information about the dataset used (Port_Mosaic) can be found in the subsection 4.2.1.

One of the most important factors is that the consecutive images have an overlap of more than 40%, but this dataset has a high frame rate and if we merge all the images, many calculations are required without adding much information to the mosaic, as the movement of the vehicle between each image is small. Therefore, it was decided to merge two images with a minimum distance of 3/4 images. This ensures that the overlap in most areas of the trajectory is greater than the minimum requirement and the images can be easily compared and combined. In Figure 48 you can therefore see an example of the collage of 2 images with a movement in the forward direction. The size of the mosaic increases accordingly. In the resulting image, you can see that the pixels are well combined. It is important to point out that this area is rich in features and even all the filters along the method are present in good numbers to produce quality for the homography calculation.

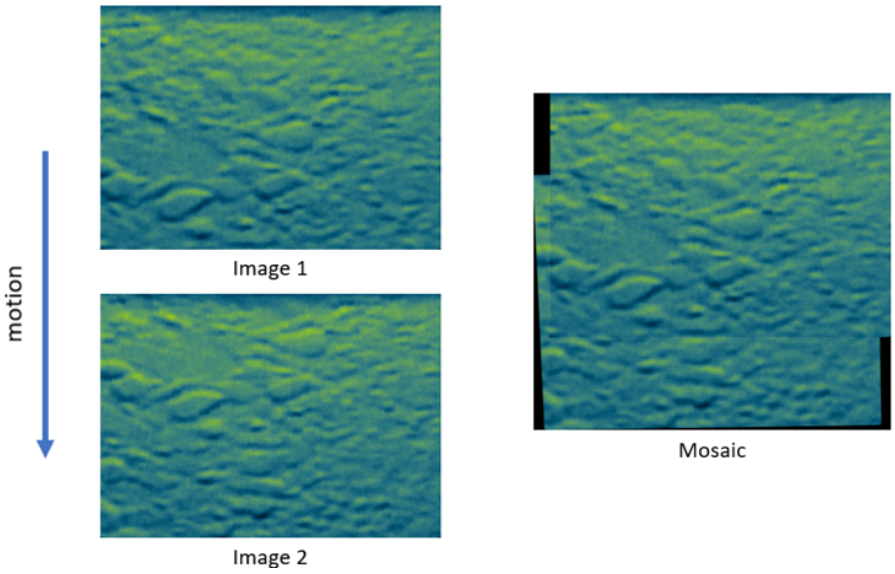


Figure 48: Example of a mosaic with two images in the stone area with a forward movement.

The following example shows the result of a mosaic with 15 images, in which on average more than 500 matches are achieved between the consecutive images. This value indicates that the result of the homography estimation is more trustworthy as it is a good value. Figure 49 (a) shows the result of the final mosaic and Figure 49 (b) shows the trajectory performed by the robot. As can be seen in the graph, the robot moves most of the time in the forward direction (about 90cm), but some deviations to the left (about 10cm) can also be observed, which can also be seen in the final mosaic and indicate the good successive warping steps.

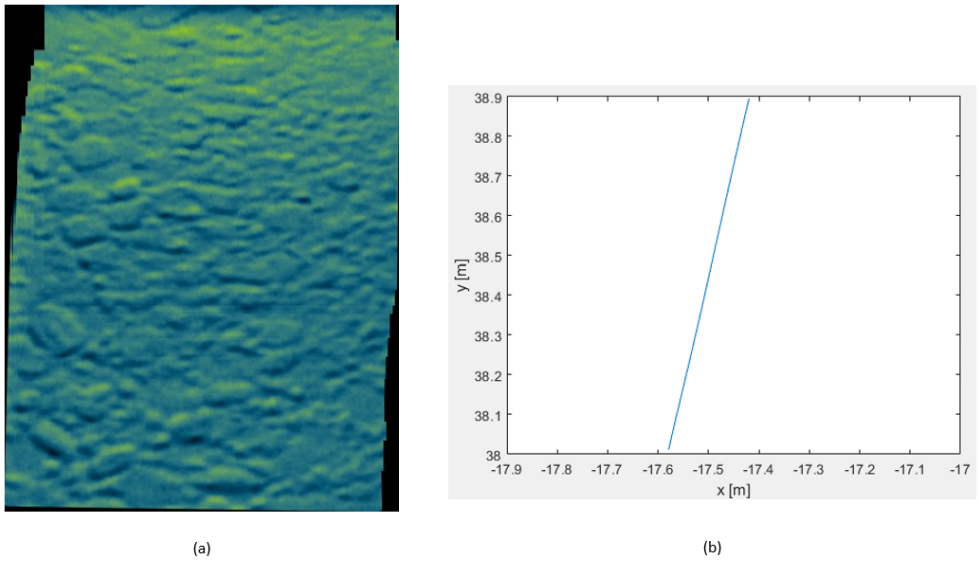


Figure 49: Combination of 15 consecutive images (a) and their movement according to the ground truth (b).

The quality and thus the number of matches used to merge the images are decisive. The higher the number of matches, the better the quality of the result, but with some configurations or in some areas it can be difficult to guarantee the quality of the mosaic. In Figure 50 you can see two examples of the same collage but with different parameterisations. In Figure 50 (a), there are only 10 good matches and the mosaic has inconsistencies due to errors in the homography estimation. Figure 50 (b), on the other hand, shows the same combination but with 29 good matches, which increases the quality of the collage. In this case, the same number of extracted features were queried, but with a higher test ratio requirement, which can greatly reduce the number of matches.

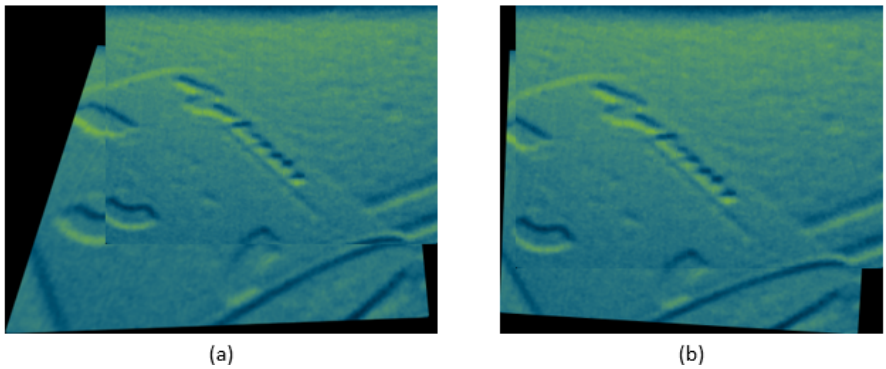


Figure 50: Example of a mosaic of 2 images with 10 good matches (a) and 29 good matches (b) caused by different configurations.

Furthermore, this type of mosaic in relation to the object is crucial to identify the type of objects that are on the seabed, as in this case where a paper cup appears in the final mosaic. However, this task is not

always easy, as the correct number of matches varies depending on the parameterisation throughout the process. One of the most important aspects is the total number of matches that the extractor should extract. Theoretically, a more reliable mosaic is created when more features are present. However, this is not always the case, because since the acoustic images are rich in textures, most of the extracted features cannot be distinguished very well from each other, which makes the process even more confusing. Therefore, it is necessary to evaluate the differences in the final result by using a different number of extracted features: 500, 1500, 3000, 4000 and 5000 per image. In the Figure 51 you can see this difference when 2 images are merged.

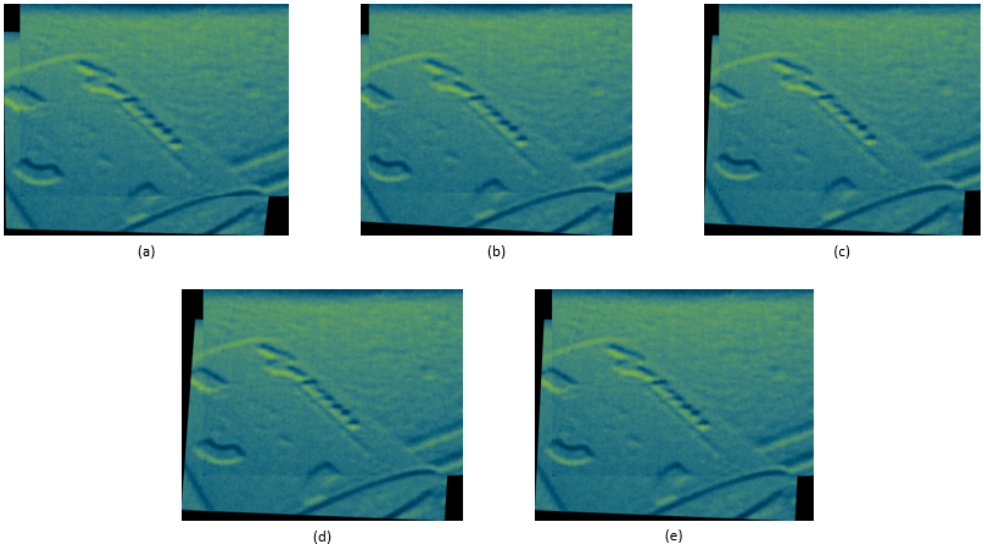


Figure 51: Differences between the merging of 2 images with: 500 (a), 1500 (b), 3000 (c), 4000 (d) and 5000 (e) extracted features.

As the visual result shows, the quality increases with the number of extracted features up to 4000 features, but in this particular case, if we use 5000 features, the quality of the results is consistent. It is also important to point out that this high number of features is not all used to match the images, as only the best matching features are used. In this case, extracting 4000 instead of 500 features in this region results in more than 250 matches with the same constraint in the test ration. Increasing the number of features in the extraction to 5000 leads to an increase of less than 10 matches in total, which is not an advantage. Therefore, extracting 4000 features is the best choice for the remaining experiments. Next, with this number of requested extracted features and a test ratio threshold of 0.8, i.e. only the features with this distance value to each other are selected, the previously selected regions (sand and grid, since the results on stones are described in Figure 48) are tested for 6 images with differences of three images between them, since the frame rate of FLS is high and yet an overlap of more than 80% between the images is ensured. The results for sand (Figure 52 (a))

and grid (Figure 52 (b)) are appealing in terms of visual quality, as the collages are well performed without significant errors or discrepancies.

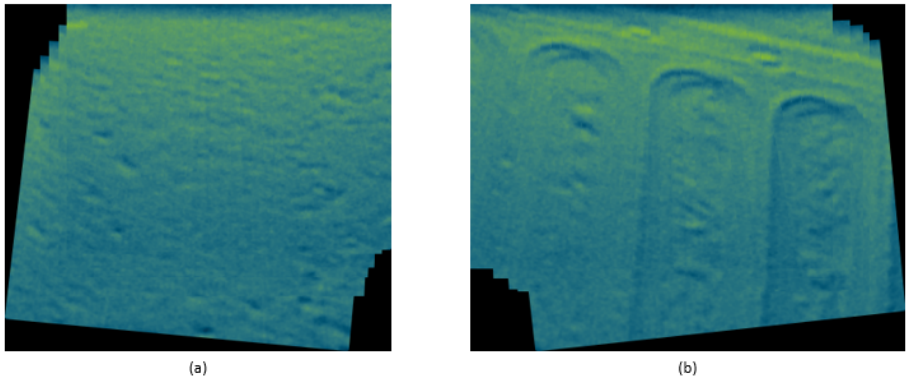


Figure 52: Final result obtained by combine 6 images of the different region: sand (a) and grid (b).

As expected, there were fewer matches in the region with sand (less 45%). The previously tested region with stones, on the other hand, always showed more than 500 matches, which corresponds to a value of more than 5 times. For this reason, the stones showed that it is easiest to combine images. The movements of the vehicle on the individual paths are clearly recognisable in the final result.

One task that is crucial in this environment is the automatic detection of objects that may appear while the vehicle is moving. If this task is performed well, the robot can take some preventive measures, such as swerving or changing trajectory. When we look at individual acoustic images, it is difficult to recognise which object it is, especially if we use a not very large FOV. For example, if we look at the individual images of Figure 53 (a), this is the case, but if we combine some images with it (Figure 53 (b)), a carabiner is visible and easy to identify. The final collage shows a good result, as the object appears complete and you can see that the seams connect the same areas in different images (not only in the metal part of the carabiner, but also in some stones that appear on the side).

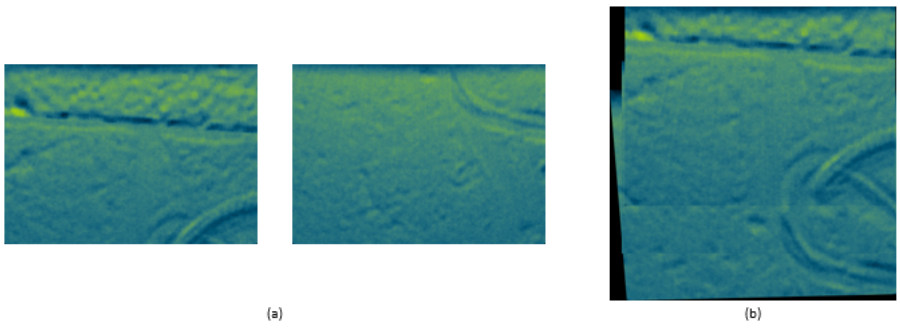


Figure 53: Individual acoustic images of a carabiner (a) and the result of the combination of successive images of this object (b).

In this way, a series of experiments with different objects that appear in the survey on the simulator with acoustic data will be carried out and analysed to identify the difficulties in this collage and some good points. First, we present two examples of the final result of the tyre obtained with successive images (see Figure 54).

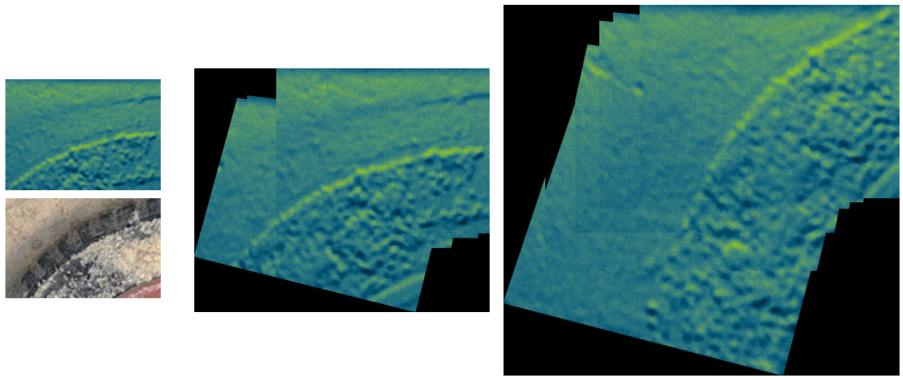


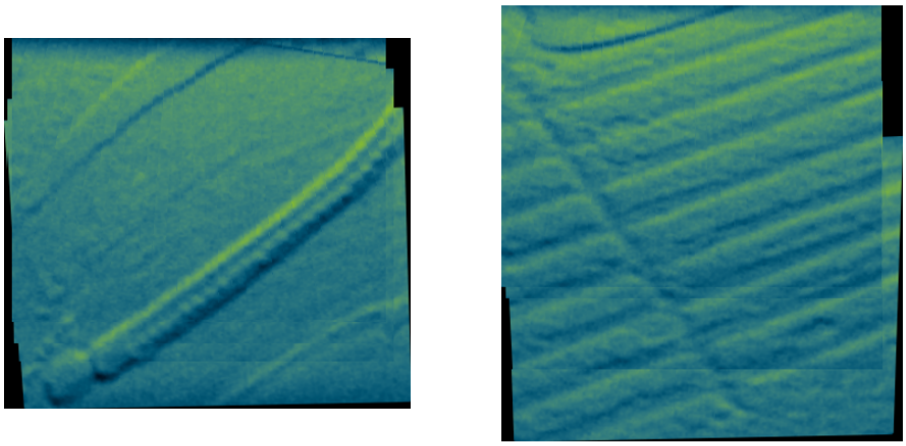
Figure 54: Individual images (acoustic and corresponding visual) of the tyre part (a) and two examples of results obtained by combining individual acoustic images of this object (b).

In Figure 54 (a) you can see an example of an acoustic image captured when navigating the robot during the trajectory and its equivalent in the image used to create the terrain of the simulated environment. As you can see, the tyre is composed of two different areas (black and red) and in the rubber part of the tyre there are some sand particles that give even more texture to this area. This area allows a good homography calculation as this textured area provides more matches between the images, ensuring an average of 1000 matches between the images. In Figure 54 (b) you can see the round shape of this object and the difference between the two different parts of the tyre. It is not possible to see the whole object because the path of the vehicle does not completely cover this part, as the vehicle deviates from its trajectory in this part of the seabed. In both results you can see the good combination between the images by the circular lines (yellow) showing the edges of the tyre.

Next, two more objects are included in the experiments: a fine grid and a rope that can appear in the harbour areas on the seabed, where other objects can also appear. This result can be seen in Figure 55.

The result of the rope (Figure 55 (a)) is visually appealing as it can be fully recognised at one end, i.e. a knot is present. The images are also well combined despite the deviations caused by the moving vehicle. It can also be seen in Figure 55 (b) that the different sections of the grid are well connected and both the vertical and horizontal metal bars have no defects. The difference between the images selected for merging is the same, but it is worth noting that the displacement between consecutive images can occur because the robot does not always have

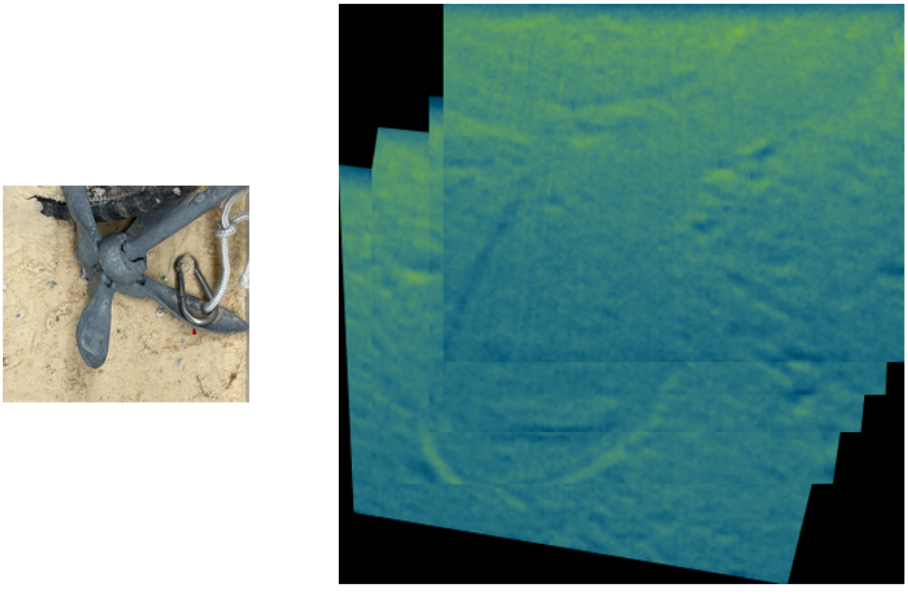
the same speed during the trajectory. The average number of matches of both regions is good and above 800, but as expected, more matches are obtained for the grid example than for the rope, since the height differences of this type of objects generate more structured data for feature extraction.



(a) (b)

Figure 55: Results obtained where a rope (a) and a fine grid (b) can be observed.

Next, a part of an anchor lying on the seabed, as shown in Figure 56 (a), is shown on the acoustic mosaic of Figure 56 (b), reconstructed with the same parameters as the previous ones, but with some inconsistencies.



(a) (b)

Figure 56: Single visual image of the anchor (a) and the result of combining consecutive acoustic images (b).

On average, only 140 matches are found. This can be explained by the region around the anchor, which consists of sand with few stones.

This number is possibly so low because some of these matches are also discarded in the next steps of the process (e.g. in the homography with the RANSAC). To try to increase this number and thus improve the quality of the final mosaic, it was therefore decided to increase the required number of features to 6000. The number of matches also increases, but only with an average of more than 25 matches, leading to the result shown in Figure 57 (a). As you can see, the collage improves its quality, but some edges of the anchor do not appear on the mosaic and some of them appear twice. Therefore, it was decided to make the test ratio more benevolent and allow more flexibility. The result in Figure 57 (b) is much more visually appealing. In this case, the number of matches increases with this change to an average of more than 700 matches between the collages, giving a more reliable result.

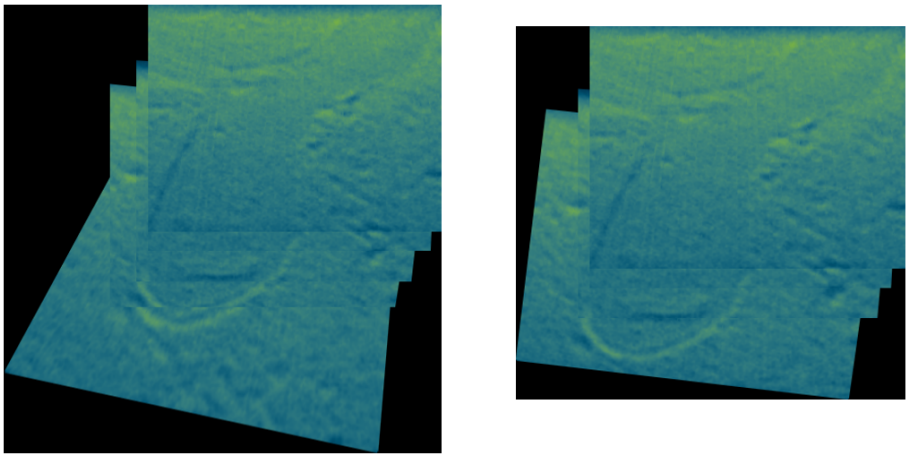


Figure 57: Differences on the obtained mosaic extracting more features (6000), but using different requirements on the test ratio: 0.8 (a) or 0.9 (b) for maximum distance.

Sometimes, however, this benevolence is not good because it can lead to incoherent results. Although objects like the plastic cup and the box have a smaller number of keypoints in the body, in most cases they have some structure points around them. The cup is integrated on the simulated seabed and the box is created in the simulator as an external object with a regular shape (parallelepiped) and without any texture on the surfaces. If you use the previously selected parameters, there will be inconsistencies in the roughness of the cup and it will be stitched crooked. Therefore, it was decided to handle the test ratio more strictly and exclude more matches (again with 0.8 for the maximum distance), leading to the result of the Figure 58 (a). In this case, there are on average only 140 matches, but with a sufficient quality to obtain good results. Otherwise, the images of the box are well combined with 0.9 or 0.8 (because it is placed above the stones and these have a high texture) and for this reason it was not necessary to use more matches if this does not lead to better results. Therefore, the result of the box

can be seen in Figure 58 (b), which gives an average of more than 380 matches and makes the images just as visible. The steams are more visible because some shadows of the original images appear on the final mosaic, but the sides of the box are well connected.

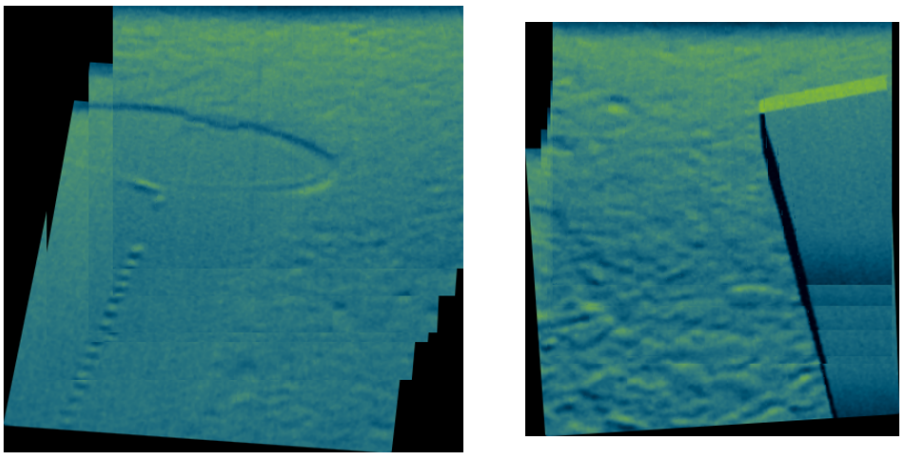


Figure 58: Result with a plastic cup (a) and a box (b) with 6 consecutive images of each object.

Another object presented in the simulated scenario was created from an upside-down vase (see Figure 59 (a)). This object has a larger shape that could not be perceived in a simulated acoustic image with the original configurations of the setup (FOV of 40°). Therefore, an incomplete representation of the vase with 3 consecutive images is obtained, as shown in Figure 59 (b), but in reality the shape of the object is not perceived or what is.

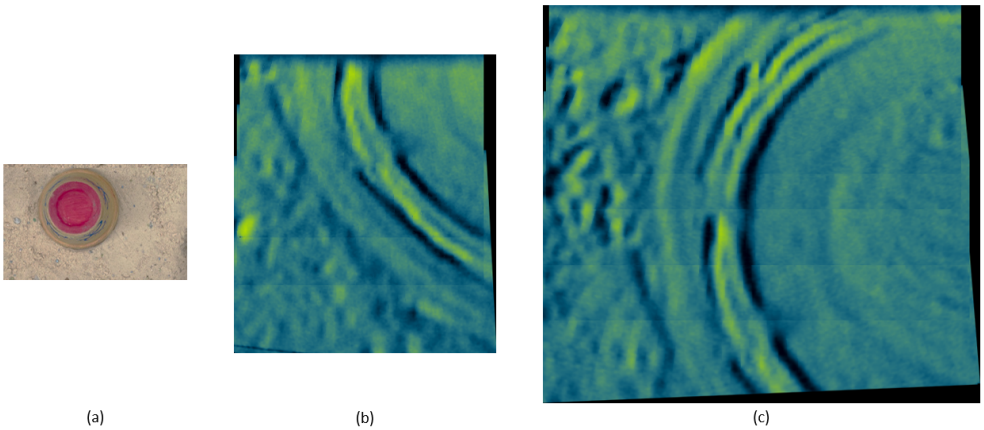


Figure 59: Single acoustic image of the vase (a), the result of combining single acoustic images with FOV of 40° and with FOV of 50°.

It could therefore be advantageous to increase the FOV to 50°. It was obvious that this small increase of the Field-of-View makes it possible to recognise a round shape of the vase, as well as the larger edge lying on the seabed. Therefore, in Figure 59 (c) the final representation of

the vase is shown, where you can see the side of the object. In addition, the central part (red in the image), which is deeper, is visible with quality on the result.

5.3.3 Discussion

2D representations, such as mosaic techniques, are important because they allow better visualisation of certain areas. When we use the images of a FLS, this is also true because the images are small if we cut off the no-echoes area to avoid the black pixels of this sensor. In the experiments, two different binary extractors were compared as they offer several advantages, such as lower memory and computation time requirements, as their representation is compact and they are used for real-time applications. Therefore, the behaviour of BRISK and ORB was compared. Based on the results obtained, it is easy to see that the number of keypoints is always lower when using BRISK, even with a high number. A high number of keypoints does not always mean a good matching result, i.e. many of these points are discarded during matching because their robustness is low. The analysis of the results shows that the ORB extractor provides a higher number of matches with the quality in the ratio test. Therefore, the most important experiments can be tested with this extractor for all cases. When using outdoor or underwater images for this task, an additional step is usually used that allows the final mosaic image to be seamless. Since the main goal of this work is to test the visual approaches in acoustic images, this step is not considered since the appearance of the objects in these images does not change over time. However, when testing the method, it was found that with large perspective changes between the images to be combined, some black areas appear in the centre of the final mosaic, as the overlapping areas are not blended. To avoid these inconsistencies, a different approach was introduced, namely a cropping step that preserves only the most important part of the mosaic without losing the idea of the movement of the vehicle. One of the most important factors for the success of the mosaic is that the consecutive images have an overlap of more than 40%. When testing the mosaic of 15 consecutive images, the motion estimated by the homography is compared with the ground truth and it is clear that it is well estimated. Finally, it is shown how important it is to have good matches, because a lower number of matches leads in many cases to an undesirable mosaic with a poor combination between the images. Therefore, the mosaics are poorly executed and often the object that appears in the image is not correctly represented (sometimes it is impossible to know what it is). In an experimental test comparing the number of required extracted features, the quality increased with the number of extracted features up to 4000 features. The results for the

different region types: Stones, Sand and Grid are appealing in terms of visual quality, as the collages were created without significant errors or discrepancies.

One task that is very important in this environment is the automatic detection of objects that may appear while the vehicle is moving. If this task is performed well, the robot can take some preventive measures, such as avoiding or changing trajectory. When we look at individual acoustic images, it is often difficult to recognise which object it is. Therefore, several images of different objects are combined in order to obtain as complete a representation as possible. For example, a carabiner, part of a tyre, a fine grid and a rope are well obtained and the final mosaics show quality in terms of visual and ease of perception of what is in the image. In some objects it is more difficult to obtain the final collage, such as an anchor, a plastic cup and a box, which have fewer keypoints. For this reason, it was necessary to increase the number of features required and in some cases change the constraints of the test ratio.

From this we can deduce the relevance of this topic and realise that it is possible to extend the findings of the visual mosaic to the acoustic images and obtain larger representations of the seabed, namely the objects, with the same assumptions, depending on the movement of the vehicle.

5.4 SUMMARY

Along this chapter, the relevance of the theme of scene reconstruction can be established, as it is crucial to interpret and understand the information about the environment. Mapping the underwater world is crucial for the exploration of this unknown environment and useful scene reconstructions are important in various fields, namely ecology, biology and geology. Therefore, a simple method was first developed to obtain a preliminary version of a 3D reconstruction of the scenes under the vehicle by using localisation with an external source (based on the results of an AruCo detection). In this way, it was possible to evaluate the performance of MSIS in this type of mapping task and observe some issues related to the elements of the robot's environment. One of the most important features of the method is that it performs two different filtering techniques: one according to the returned intensity and another according to the scan angle (which can be switched off). In addition, unlike the usual techniques, only the central part of the 35° arc of the sensor is considered occupied (about 20°) in order to reduce memory consumption when creating the map without losing important information. Most of the materials tested are clearly visible with 40% gain. However, in order to choose a single value for all

materials, a value of 50% gain was used, which proved to be the most suitable. If only the centre part of the arc is occupied, the constant error in the measurement dimension of the direction of movement is reduced to 33cm. The method makes it possible to map smaller objects with straight lines without having to resort to the scan angle filter with a maximum error of 5 cm, which is good for real-world applications. It is possible to observe objects with more complex shapes and create maps with trajectories with directional changes. The z dimensions of the objects are correct and the MSIS proves to be effective in detecting obstacles near the sensor. Finally, performing a discretisation process with different step sizes shows that it is possible to achieve a simpler representation of the scene, capable of collecting some information about the objects, with lower storage costs.

In the second part of this chapter, the mosaic technique is investigated with acoustic images from a FLS that allow a better visualisation of certain areas. Since the assumptions used in this work are the same as for visual mosaicking, some visual extractors are tested to find out which are the best. When analysing the comparison results, it can be seen that the ORB extractor provides a higher number of matches with quality. All tests were performed with this extractor and under the assumption that the images to be used for the mosaic overlap by at least 40%. Overall, the result of merging the images is good. The motion estimated by the homography is compared with the ground truth and it is clear that it is well estimated. When comparing the number of required extracted features, it is seen that the quality increases with the number of extracted features up to 4000 features, so this was the best choice for the remaining experiments. The results for the different regions such as the stones, the sand and the grid are appealing in terms of visual quality as the collages were created without any significant errors or discrepancies. One task that is very important in this environment is the automatic detection of objects that might appear during the exploration. For objects such as a carabiner, part of a tyre, a fine grid and a rope, the results of the mosaic are well combined and the final mosaics show quality in terms of visual and ease of perception of what is seen in the image. For some objects such as an anchor, a plastic cup and a box, which have fewer key points, it was necessary to make more parameterisations to increase the number of matches obtained and at the same time the quality.

The reconstructions obtained using acoustics enable us to better understand the underwater scenarios when the raw data is available and we process it using these methods. At the same time, it provides the robot with a better interpretation of its environment, which can be integrated into other applications.

6 VISUAL OBJECT RECOGNITION

This chapter presents some experiments and analyses on object recognition and its semantic segmentation. This allows you to check how some methods and approaches on this topic work in underwater scenarios.

The chapter is organised as follows: section 6.1 introduces some applications and concepts related to the topic of visual object recognition and its segmentation by visual images. Section 6.2 contains a preliminary study on critical object recognition, comparing an in-house implementation with one from the literature, using a previously developed dataset. In section 6.3, the concept of object recognition is extended to semantic segmentation in an image of objects. For this purpose, a comparative study of different literature approaches is applied to a known underwater dataset. Then, in section 6.4 the investigation of some strategies to improve the previously obtained results is shown in order to apply semantic segmentation of images in an underwater environment. Finally, the main conclusions of this chapter are presented in section 6.5.

6.1 INTRODUCTION

Object recognition has become an attractive field due to its close connection with video analysis and image understanding. The main goal is to enable machines to interpret and understand visual information in a similar way to humans. As ocean exploration is far from complete, the development of recognition systems that can recognise objects in different conditions is crucial. The reason is that this environment is unknown and there are some critical and dangerous objects in the sea, such as mines or chains, whose automatic recognition is very important for various tasks. In recent years, many of these tasks have been taken over by robotic vehicles, which therefore have to adapt their behaviour to the objects they find. There are several sensors that can help in object detection, but the visual sensors have obvious advantages, such as some already described, e.g. their cost, their good performance in detecting nearby structures and the amount of information provided. However, due to some phenomena related to the propagation of light in water, their use in this environment is limited as detection requires boundary detection, which is a challenge in this environment. The turbidity of the water, the different brightness ratios on the images

and the background confusion caused by reef features and underwater plants make the process of image classification a difficult task. Recently, deep learning has emerged as a powerful tool to overcome the shortcomings of conventional image classification methods, especially in this environment. Deep learning therefore provides a way to extract unique, invariant and robust features that can withstand the distortions and variations in images. However, the output of these methods is only this bounding box with the object class and the coordinates of the object.

Nowadays, however, more and more diverse tasks can be performed by autonomous vehicles instead of endangering humans during underwater exploration. Therefore, over time, a higher level of intelligence has been observed in underwater robots than before. For this autonomy in exploring this unknown environment, it is becoming increasingly important to enable the robot to decide for itself how to act or react when it sees a particular object. This requires a good semantic map on which the robot can map the structure of the scenario and at the same time recognise the objects and structures involved. This therefore requires an object recognition system that can not only identify an object, but also its shape and other features that allow a better interpretation of the scene. Recently, the term object segmentation has become established, which goes one step further than object recognition. This involves not only the identification and localisation of objects, but also the precise delineation of individual objects at pixel level in an image, i.e. dense classification. The result is a high-resolution mask that provides a detailed visualisation of the shape and boundaries of each object and its class. In addition to autonomous navigation, semantic segmentation can also be used in the search for archaeological artefacts or even in the inspection of underwater structures such as platforms, cables or pipelines. Other applications can be found in marine biology and ecology in the identification of species, but also in the search and rescue of missing persons or in military defence through the classification of mines. It is therefore crucial that the results obtained have a higher degree of accuracy and that the result provided by the segmentation has as few errors as possible so that the robot can understand the entire scenario well and make appropriate decisions when necessary.

6.2 CRITICAL OBJECT CLASSIFICATION

Recognising objects in underwater environments is a particular challenge compared to recognising objects out of water, but it can be an important tool. It combines the challenges of the environment with

underwater imaging and the idea is to develop and deploy the most robust algorithms and technologies to overcome these limitations.

Therefore, it is necessary to test and develop methods that solve this problem in detecting some critical objects under different conditions and characteristics. To do this, it is important to understand the main steps of this process and what factors can improve or impair the good results. The process of object classification consists of three main phases: data collection, training and evaluation. Inspection and monitoring tasks on the seabed are increasingly performed by autonomous robotic vehicles and the behaviour of these platforms must be different for each object. For example, if the robot finds a mine, this is dangerous and the robot must perform a pre-programmed action (escape or destruction). However, if the robot sees a pipeline or a chain, it can bypass the object and continue the mission. It is therefore important that the robots recognise different objects with high accuracy. One of the most important steps in solving a classification problem in the real world is obtaining the data for the training process. Thus, the quality of the input data is the most important factor for the accuracy of the results. Since there is a lack of underwater data with the intended critical objects, it is important to create a dataset that can be used to test this task.

The first step in this study was therefore to develop a dataset with a series of images of various man-made objects that are considered critical in underwater operations. The *CriticalObjectDataset*, which is well explained in section 4.4.1, therefore contains 7 classes: Anchors, Chains, Pipelines, Fluctuation Modules, Mines, Buoys and 3D Markers (see Figure 60).

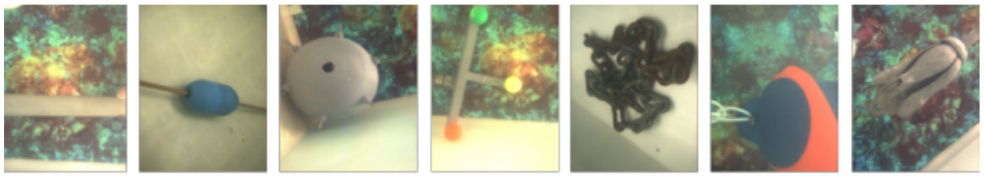


Figure 60: The 7 critical objects used in the *CriticalObjectDataset*.

6.2.1 Methodology

In many real-world cases today, it is rarely necessary to develop new architectures for object recognition, as there are different approaches that can be used to solve the problems. However, sometimes it is important to develop methods that rely on a Convolutional Neural Network (CNN) to evaluate their performance and computational resources in other contexts and under other conditions. In networks that use only fully connected layers, all nodes in one layer are fully connected to all nodes in the previous layer, creating a complex model.

This complexity is bad for the training phase and often brings no additional benefit (many features are localised). Otherwise, CNN applies small filters so that the number of training parameters is tiny and it can use many layers simultaneously. However, it is important to note that the depth of the network increases the complexity of the model and the ideal is to find a compromise between complexity and performance. Therefore, in a preliminary study, a network architecture was developed that contains a combination of different layers, as can be seen in Figure 61.

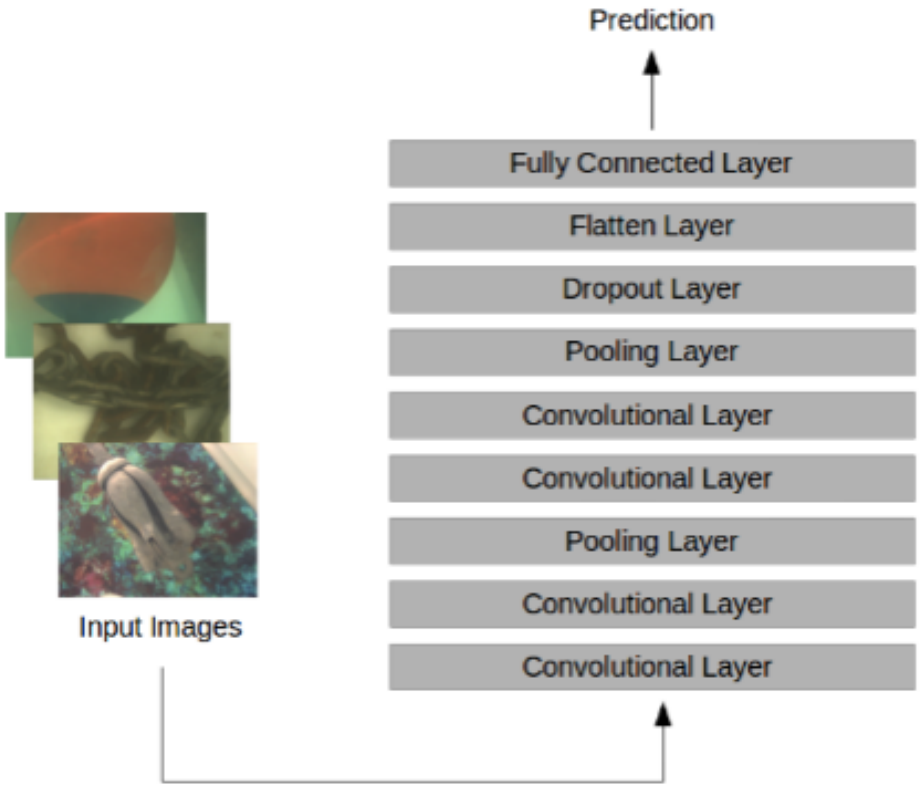


Figure 61: Process overview of the proposed CNN method.

The main purpose of the convolutional layers is to extract features from the input image while preserving the spatial relationship between the pixels. To reduce the size of the features and to ensure that the algorithm does not focus on small changes in orientation and position, the pooling layers were added. A dropout layer was used to improve the generalisation ability of the model and prevent overfitting of the final results. In many cases, a CNN has to deal with the problem of degradation, i.e. when the training error increases with depth. On the other hand, deeper neural networks are more difficult to train. For these reasons, the Residual Neural Network was introduced to facilitate the training of the network [86]. This architecture makes it possible to overcome the challenges of training very deep neural networks. It has proven to be easy to optimise and has so far achieved better results than the other deep methods in the literature. Therefore,

for comparison with the results of the developed CNN, the ResNet-18 was selected, a specific variant with a total of 18 weighting layers. This framework also allows the use of weights trained with the ImageNet, a large visual database developed for use in visual object recognition research, to start the model, which usually improves the accuracy of the final results.

6.2.2 Experimental Results

A series of experiments were conducted with different splits of the developed dataset. Therefore, the general performance of the in-house algorithm is demonstrated using a validation set and a comparative analysis with the state-of-the-art method (ResNet) is presented. In this way, four evaluation parameters were determined for each study case:

- Training Accuracy (TA) - this is a measure of how well a machine learning model performs on the training dataset.
- Validation Accuracy (VA) - it indicates how well the model generalises to new images that it has not seen during training.
- Test Accuracy (TeA) - this provides an estimate of the model's performance in real-world scenarios.
- Processing Time (PT) - this gives the time taken to train a model on a particular dataset

In order to achieve good generalisation and good results in predicting the object classes when using the models, it is important to monitor all three metrics. In this way, it is possible to understand how well a machine learning model works in real-life scenarios. The evaluation of the proposed CNN is therefore performed with an initial set of images randomly selected with an appropriate algorithm - *DatasetA*. The only aspect that was checked by the user was the presence of different perspectives between the training set and the validation set to avoid overfitting. The main features of this dataset are: the training and validation sets consist of a mixture of images (pure and turbid water), each class is divided by an 80/20 ratio for training/validation, the training images are 3525 and the validation images are 967.

Figure 62 shows the performance obtained in both phases: training and validation as well as the loss values obtained in the training phase for 50 epochs. The VA obtained in the best epoch was 76%, which means that the algorithm found more or less 735 images correctly. Otherwise, Figure 63 illustrates the behaviour of the object classification along the different epochs.

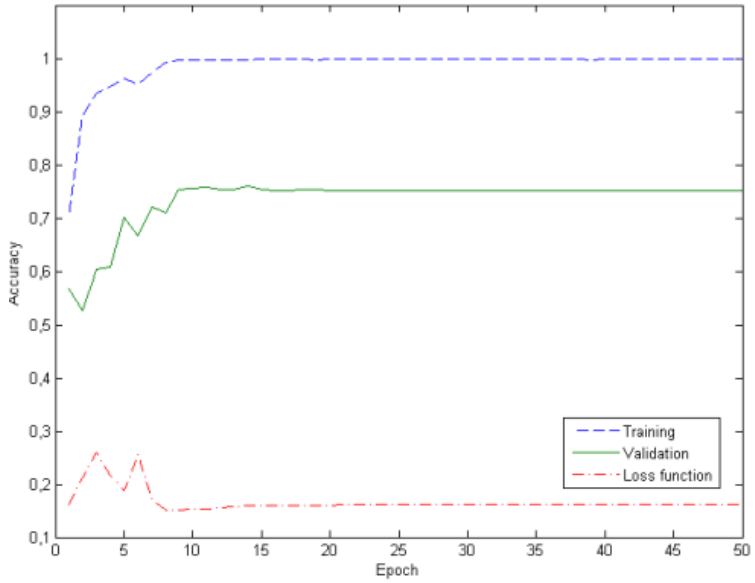


Figure 62: The performance achieved in both phases: training and validation, as well as the loss value during 50 epochs.

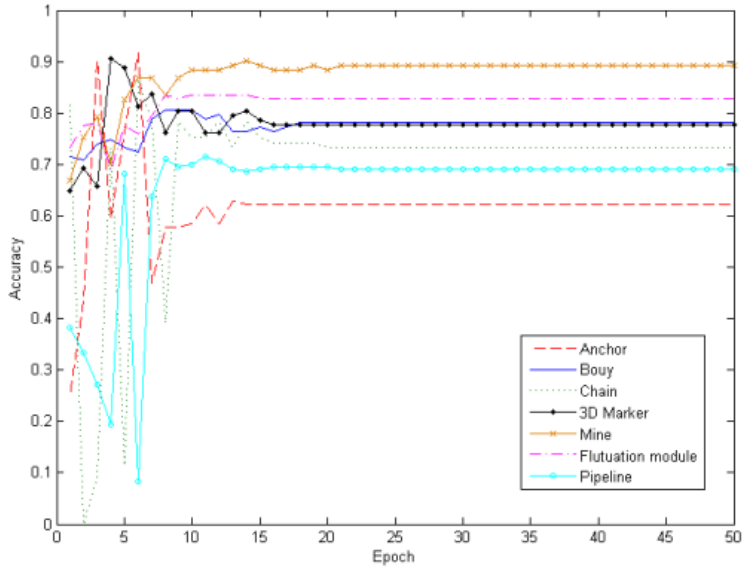


Figure 63: Achieved performance in terms of accuracy for each class over 50 epochs.

It can be seen that anchors and pipelines are the most difficult classes to classify. This can be explained by their grey colour, which in some situations can be confused with the background. Conversely, the mine is the object with the higher accuracy, which helps to improve the final accuracy achieved. Although it also has a grey colour, it has a characteristic shape. To validate the results obtained, the dataset was validated with the algorithm ResNet without pre-trained weights (ResNet_v1), see Table 11. In terms of learning conditions, this ResNet

variant is the only one that can be directly compared with the proposed approach, as it does not contain any prior knowledge about the objects.

Table 11: Accuracy metrics and time required when using the proposed method and ResNet_v1 for the *datasetA*.

DatasetA	In-House	ResNet_v1
TA	99 %	99 %
VA	76 %	86 %
PT	37m22s	89m25s

It can be concluded that the results obtained with the proposed approach are not very far from the results of ResNet. Moreover, the ResNet method requires more than twice of PT. This fact therefore has a significant impact on real-time robot applications. As expected, the anchors and pipelines were the classes that performed the worst. To evaluate the performance of the proposed algorithm without these classes, the training and validation phases are performed again, but now with only 5 object classes. In this case, the performance of the developed method increases to a VA of 84% and the ResNet to a VA of 96% (about 10% in both cases).

To evaluate the robustness of the proposed approach to new conditions, a new dataset (*datasetB*) was created to complicate the classification process. Thus, it was possible to train the model only with images acquired under ideal conditions (pure water, without suspensions) and test it in turbid water to simulate the real conditions. It is therefore to be expected that the performance will decrease significantly. The ratio of training and validation images was maintained. Table 12 shows the comparative results between the proposed method and three ResNet variants: ResNet without pre-trained weights (ResNet_v1), ResNet with pre-trained weights but only training the last layer - Fully Connected (ResNet_v2) and ResNet with pre-trained weights (ResNet_v3). As expected, the Validation Accuracy decreases sharply in all cases except for ResNet_v3, which has a VA of 94%.

Table 12: Accuracy metrics and time required using the proposed method and ResNet variations for the *datasetB*.

DatasetB	In-House	ResNet_v1	ResNet_v2	ResNet_v3
TA	81%	98%	99%	99%
VA	25%	47%	30%	94%

In this case, the fact that the training is performed in all layers made it possible to learn new features about the different objects during the successive epochs. In addition, the use of pre-trained weights helps in the first step, as some interesting features such as

boundaries and shapes (which may be similar to the objects in the developed dataset) are included. This fact is also reflected in the results of ResNet_v2 (VA of 30%), which, as expected, has the worst accuracy as only the last layer is trained. This approach therefore does not learn any new features during the training phase. However, the initial values showed a certain knowledge of the classes, which enabled the accuracy achieved. So far, the developed approach has only been validated. However, a crucial step is the prediction using the best model obtained from the best epoch during the validation phase. Therefore, a new dataset (called *datasetC*) was created with randomly selected training images so that all conditions were considered. For the test set, only images in turbid water were used to simulate the real test conditions, and the data were split as follows: 70%, 20% and 10% for the training, validation and test phases, respectively. See Table 13 to observe the results obtained. The developed method maintained the TeA achieved in the validation phase (about 75%). In contrast, the ResNet_v1 (without pre-trained weights) showed good accuracy in the validation phase, but its accuracy decreased by 15% in the test phase. This situation can be explained by the difficulty of the test images.

Table 13: Accuracy metrics and time required when using the proposed method and ResNet_v1 for the *datasetC*.

DatasetC	In-House	ResNet_v1
TA	87%	98%
VA	75%	91%
TeA	74%	76%

Finally, the *datasetD* was created with user-selected images to ensure that the majority of images in the test dataset matched the simulated real-world conditions. An attempt was made to make the training dataset as heterogeneous as possible in order to correctly learn the main features of the objects under different conditions. In addition, the validation and test datasets were selected with different viewpoints to avoid overfitting. Table 14 shows the behaviour of the proposed method and ResNet_v1.

Table 14: Accuracy metrics and time required when using the proposed method and ResNet_v1 for the *datasetD*.

DatasetD	In-House	ResNet_v1
TA	98%	98%
VA	70%	87%
TeA	67%	92%

It can be observed that the performance of the developed implementation has decreased. This can be explained by the difficult requirements in the test set. Otherwise, ResNet_v1 has a higher validation and testing accuracy. Its learning ability during the training phase is therefore remarkable, which is reflected in the Test Accuracy (92%).

In addition, two further study cases were performed:

1. Images with white background (clear and turbid water) were used in the training phase and the tests were performed with reef background (clear and turbid water). The results show the influence of the training phase on the classification process, as the validation accuracy decreased significantly for both methods: the developed approach showed a TeA of 20% and ResNet_v1 a TeA of 37%, which means a decrease of 50% in both cases.

2. The training was performed with only 6 objects and 7 classes were used as input for the test process. In this case, the developed approach achieved a performance of 51% and the ResNet_v1 reached 64%. Therefore, it is crucial that the approaches contain mechanisms that can derive a new object in the testing phase (which was not considered during training).

6.2.3 Discussion

From the results obtained, it can be concluded that the performance of the developed method is good enough to correctly recognise the objects with an accuracy of 75% under most conditions. Although all classes can be classified with the developed method, there are some classes where this task is more difficult. This fact can be explained by the colour, which can be confused with the background (reef). The object classification without these classes (anchor and pipeline) improves to a VA of 84%. The results obtained with the ResNet without pre-trained weights (ResNet_v1) can be compared with the developed method, since in both cases the weights are initialised randomly. The performance of this ResNet variant is better than that of the developed method, but it involves higher computational costs and requires more processing time, which is sometimes very concerning, especially for real-time operations. ResNet with pre-trained weights (ResNet_v3) offers higher performance. Otherwise, the results of ResNet_v2 prove that the initial weights are very important during the classification process, as they can reach 30% of VA even without training the convolutional layers. When evaluating the robustness, the results have shown that both the training under ideal conditions and the validation and testing under the simulated real conditions do not work correctly as expected. In general, classification underwater is difficult, especially in turbid water as it is most similar to the real

environment - it leads to blurred boundaries of the objects. In this extreme situation, it makes more sense to train the model with the ResNet with pre-trained weights (ResNet_v3) when computational cost and PT are not an issue. In the final experiment, it was found that the developed method can be used in many situations in a preliminary form, since the training was performed with heterogeneous images. In this way, the model learns the features that can be used to better distinguish the objects. Good accuracy of classification methods is often accompanied by high computational complexity, which requires the use of GPU.

As mentioned above, the main goal of the task described in this study is to identify and localise objects within an image or a video by merely detecting the presence of objects in an image. The output of these methods is therefore only a bounding box with the object class and the coordinates of the object. Nowadays, however, robots increasingly require a good semantic map and thus a representation of the scenario. This therefore requires an object recognition system that can not only identify an object, but also its shape and other features that allow a better interpretation of the scene. Recently, the concept of object segmentation has become established, which goes one step further than object recognition.

6.3 COMPARATIVE STUDY OF SEMANTIC SEGMENTATION METHODS

Image segmentation is therefore the process responsible for dividing an image into several regions or objects. If this task is performed successfully, the robot can estimate the geometry of the scene and draw some conclusions about the interaction of the objects. In recent years, this problem has been studied in the literature due to its importance, but more in outdoor environments than in underwater environments. In most cases, this segmentation is done in images (visual or acoustic), as this is the best way to obtain the most information about the objects and their features. In this case, the goal is to use the visual sensors, but sometimes the images in the underwater environment are noisier and more blurred, so it is more difficult to recognise the boundaries of the objects, which are crucial for this task. Some of the traditional scene segmentation methods already developed are applied in underwater environments, achieving correct segmentation only in certain situations and relying on heuristics. Inspired by the success of deep learning architectures, some approaches have resorted to machine learning to learn the key features of each class and thus develop a generalised solution based on a reliable set of training images. To this end,

datasets from underwater environments must be used to train these methods and evaluate their behaviour in this environment. However, the larger datasets that exist in the context of semantic segmentation have been obtained in outdoor scenarios. When suitable underwater datasets do exist, they are sparse in terms of perspectives for each class, poor in terms of diversity of classes, or contain few samples for each class, making it difficult to obtain good visual segmentation results. The poor image quality of the recorded data, which in most cases does not contain true colour information due to light attenuation, noise and interference, makes this task challenging. Therefore, the idea of this study is to select a more suitable segmentation method from the underwater imaging literature to apply these techniques in harbour areas where there are many interesting objects and structures to explore. For this end, this study uses a suitable dataset available online, already described in section 3.4, to verify the quality of the results obtained with the class segmentation models. The main purpose of this study is to understand where the main difficulties lie and how performance can be improved if necessary. However, these decisions are not always trivial as many aspects need to be considered to find the best solution for the intended application. Hence, a comprehensive comparison on the performance of some machine learning based literature approaches for image segmentation that have been studied mainly in outdoor domains.

6.3.1 Methodology

In a conceptual context, semantic segmentation can be viewed as pixel-level classification, meaning that each pixel in an image is associated with a class label. Therefore, in the world of machine learning, many approaches have been developed for application in an outdoor environment, but they also need to be applied and tested in other environments, namely underwater scenarios, as this can be very important and challenging for the development of underwater robotics. Therefore, the first step after reviewing the implementations of the different semantic segmentation methods is to compare them and see how well they can be applied in an underwater context. Factors such as learning rate, computational effort and time requirements should be taken into account, because even though data processing is becoming more and more advanced nowadays, there are always limits to the vehicles. For this reason, and because the main goal is to compare some pre-developed approaches, a repository implementing common semantic segmentation architectures was used [87]. This repository also compiles the data loaders of 7 known datasets for outdoor scenarios and also allows the application of some augmentation techniques that increase the number of samples of each class, which is crucial in

underwater scenarios because, as mentioned before, data collection is more difficult. To this end, it relies on *PyTorch*, an open-source machine learning framework based on the Python programming language, and the Torch library, which is commonly used to build deep neural networks.

In a previous work with a difficult outdoor scenario, from which some conclusions can be drawn about the results of challenge scenarios, all implemented architectures are compared. Based on these results and the knowledge of the current state of the art in this field, only the three best performing ones are selected for this test in an underwater scenario: Fully Convolutional Networks (FCNN) [88] - with two different versions: 32 and 16 bit, Pyramid Scene Parsing Network (PSPNET) [89] and SegNet [53]. The FCNN is a rich class of modern classification models. The authors extend these classification networks to segmentation, improve the architecture by combining layers with multiple resolutions, simplify and refine the spatial precision of the output, and accelerate learning and inference. They use only locally connected layers and avoid the use of dense layers, which means fewer parameters (allowing the networks to be trained faster), and train pixel by pixel. The PSPNET is an effective Pyramid Scene Parsing Network for complex scene understanding, i.e. a superior framework for pixel-level prediction tasks. The pooling function of the global pyramid provides additional contextual information, and this pyramid enables the acquisition of further layers of information. It effectively captures contextual information at different scales, enabling accurate and detailed segmentation of objects in complex scenes. The pyramid pooling module contributes to its success in processing objects at different scales in one image. Finally, SegNet is a deep, fully convolutional neural network architecture for semantic segmentation at the pixel level. This trainable segmentation method consists of an encoding network, a corresponding decoding network and a subsequent pixel-wise classification layer. It was developed with the main goal of developing applications for scene understanding. The use of max-pooling indices in decoding is a characteristic feature of SegNet that contributes to its ability to reconstruct detailed segmentation maps and is designed to be efficient both in terms of memory and computation time during inference.

Evaluation of the methods

The visual results of image segmentation are the best way to observe the results, but when we work with different classes and images, it is very difficult to verify them all and quantify the results.

For this reason, it is crucial to analyse the accuracy in classifying objects with a correct class, the number of pixels correctly assigned as true pixels for a class, the ratio between the accuracy and the time

needed to train the image set to achieve this accuracy, etc. Therefore, in addition to the visual information provided by the segmentation predictors, the results of the different semantic segmentation algorithms can be evaluated using appropriate measurements to derive important conclusions. The repository used and briefly described above also implements several methods to measure the accuracy of the segmented data. These measures are:

- **Overall Accuracy** - it refers to the percentage of pixels in the image that have been correctly classified. This metric is not particularly suitable if we use an unbalanced dataset, i.e. with classes whose elements occur much more frequently on the test image than others. Otherwise, this metric can also lead to misleading results if the class representation within the image is small.

$$\text{Overall Accuracy} = \frac{\text{Number of correctly classified pixels}}{\text{Total number of pixels}}$$

- **Mean Accuracy** - this metric takes into account the object class of each pixel if the pixels are correctly classified and is therefore a global measure. It is a pixel-wise accuracy measure that calculates the average accuracy of the correctly classified pixels across all classes. It therefore takes into account the different sizes of the objects to be classified, as there are classes as small as some fish or as large as wrecks. It is therefore more suitable for evaluating unbalanced datasets than the previous measure.

$$\text{Mean Accuracy} = \frac{\sum_{i=1}^{N} \text{Correctly classified pixels for class } i}{\sum_{i=1}^{N} \text{Total pixels for class } i}$$

  where:

  - N is the number of classes.
  - $\sum_{i=1}^{N}$ Correctly classified pixels for class i is the sum of pixels that are correctly classified for each class.
  - $\sum_{i=1}^{N}$ Total pixels for class i is the total number of pixels for each class.

- **Mean Intersection-Over-Union** - it measures the mean value of IoU in the different classes. It is therefore one of the best metrics to reflect the performance in the segmentation processes as it calculates the ratio between the area of overlap between the prediction and the real observation with the area of union of both.

$$\text{Mean IoU} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{True Positive}_i}{\text{True Positive}_i + \text{False Positive}_i + \text{False Negative}_i}\right)$$

where:

- N is the number of classes.
- True Positive_i is the number of pixels that are correctly classified as class i.
- False Positive_i is the number of pixels that are incorrectly classified as class i but are not actually class i in the ground truth.
- False Negative_i is the number of pixels that are actually class i in the ground truth but are incorrectly classified as a different class.

Therefore, a comparative analysis between the methods can be used to determine which method has higher performance values. In this way, it can be deduced which semantic segmentation algorithm is best suited for the context and conditions at hand. It is interesting to achieve better results for most metrics, i.e. more than possible. It is also important to find a compromise between good results and time, but time is not the most important thing in this context as the training should be done before the real operations. So it is important not to waste time unnecessarily (i.e. without an increase in performance or other benefit), but the idea is that the training process is as general as possible so that the method can learn all the intended classes well and segment them in the test image.

As mentioned above, visual data collection in this world is not easy, as it is an expensive task that requires appropriate materials and communication platforms and is often dangerous for divers. Moreover, this task brings several problems, such as light absorption, turbidity and limited range, which reduces the number of good quality images. However, for these methods to work well in this type of task, which requires the training of models, suitable datasets must be used. However, one of the main problems with visual methods in this area is often the lack of complete datasets or multiple images that can be considered for different contexts or classes. If there are suitable datasets for the general underwater context, they often do not have ground truth values, which is a challenge if they need to be labelled manually, or they are often sparse in terms of classes (often a dataset contains only one class, e.g. for fish classification) and have a smaller number of samples (many of them contain a total of 1000/2000 images, which is considered a small dataset). In order to use an autonomous vehicle to monitor different areas, such as harbour areas, the vehicle must be able

to accurately identify the objects present in the environment in order to decide how to react to the object. This would also allow the robot to create a semantic map that provides it with a more comprehensive representation of the environment. For this purpose, it is necessary to use a general dataset with different objects to investigate the application of segmentation methods for these areas. When analysing the existing datasets in the literature, the SUIM dataset proves to be the most attractive as it contains a large number of classes that may be common in the underwater world. Thus, the dataset consists of 1525 images for training and 110 additional images that can be used to evaluate the performance of the algorithm. These images were collected from different domains and some of them belong to other datasets and were selected by the authors with the aim of making the dataset generic. This dataset is a challenging dataset and includes eight classes that are very different and have different characteristics such as size and colour perspective. Some examples of images of the dataset and the corresponding masks with the regions of the individual labelled objects can be found in Figure 64. Regarding the size of the image, there are a number of sizes that need to be standardised before the testing process to allow for a consistent evaluation of the results. Therefore, the size of the image was normalised before model training, as is common in computer vision.

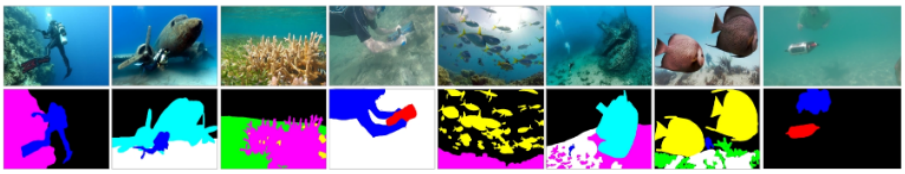


Figure 64: SUIM dataset: real images and the corresponding ground truth masks. Adapted by [79].

There are 8 classes: Waterbody (Class 0), Human Divers (Class 1), Robots or Instruments (Class 2), Reefs and Invertebrates (Class 3), Plants (Class 4), Wrecks or Ruins (Class 5), Fish and Vertebrates (Class 6), and Seafloor and Rocks (Class 7). However, a key feature of this dataset is that the number of samples for each class is unbalanced, which complicates the main task of object segmentation, as the class with the highest representativeness is represented by 80% and the class with the lowest representativeness by less than 10%. Figure 65 shows the representativeness of the classes in the original dataset, i.e. in how many images a class is present, according to the respective colour that appears on the masks. Unfortunately, this is a common problem in obtaining images for this purpose, as the images of each class are taken in different perspectives, environments and lighting conditions, but never in the same number. This problem affects the accuracy of some classes, i.e. the worst results in terms of accuracy are obtained in

the classes where the number of images is lower. It is also important to note that the classes do not appear in isolation, i.e. in each image there are several classes (in some images 5/6 classes appear), which makes it difficult to solve this imbalance.

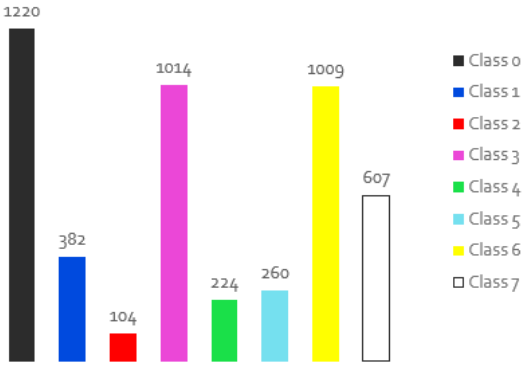


Figure 65: Distribution of the individual classes for the images of the original training set.

It is also worth noting that the images selected for the test are difficult as they have some noise, multiple classes in the same image and a high degree of colour attenuation, as can be seen in the Figure 66.



Figure 66: Difficult images to test the algorithms: noise, light attenuation and many classes per image.

6.3.2 Experimental Results

Since the goal is to compare some previously developed approaches, some configurations can be applied in terms of training configuration, such as the number of workers, the number of iterations, the loss, as well as the configuration of the optimiser and the learning rate. Therefore, it was necessary to perform a series of experiments with the different constraints and analyse the performance of the different methods capable of classifying the objects and segmenting them in the

visual data, with the main objective of selecting the most appropriate method for the intended context. In the first experiment, the four implementations are tested and compared to determine the most suitable model for the segmentation of objects in a purely visual approach. The parameters of the training process, such as the image size, the optimiser (Adam) and the loss method (multiscale cross entropy) applied to the four models are the same in order to determine the model with the best performance in a feasible way. These parameters were selected based on experience and information from the literature as well as some results from previous work in challenging scenarios. Therefore, the metrics described above are used and an individual analysis is performed to compare the IoU for each class and to determine some facts about the difficulty of each object. The training process for this test was performed with 100000 images, with the validation process performed after every 1500 iterations, resulting in a total of 66 validations. The total time of the process and the time spent on processing per image are presented as the aim is to achieve the best performance in the shortest possible time, although this is not the most important factor in this case. Table 15 shows the comparative accuracy results of the four implementations in the final validation.

Table 15: Accuracy [%], IoU [%] and time results after 100000 training iterations.

	Accuracy		IoU	Time	
	Overall	**Mean**	**Mean**	**Total**	**Mean**
FCNN_16s	76.7	55.3	44.8	12h21m	0.06s
FCNN_32s	76.6	55.3	43.2	12h20m	0.06s
SegNet	79.6	62.7	50.6	09h08m	0.02s
PSPNET	78.4	57.5	46.6	18h58m	0.15s

As you can see, in terms of overall accuracy, the four implementations provide a useful value for use in real-world environments, exceeding 75% of performance, but the best results are obtained by PSPNET and SegNet. In terms of time, the data returned is more varied and this factor needs to be analysed. The differences in training time are large and it can be concluded that the PSPNET implementation, for example, requires a lot of time and therefore must be a problem in practise. Otherwise, the other implementations require less than 7 seconds for the image, which is a more pleasant value. Even though the results at the end of the training process are quantitatively similar, the image segmentations obtained for the test images after 30 evaluations (in the middle of the process) were created to better understand the quality of the learning process. Figure 67 shows the segmentations obtained with PSPNET, SegNet and FCNN_16s (the results are the same for both versions of FCNN).

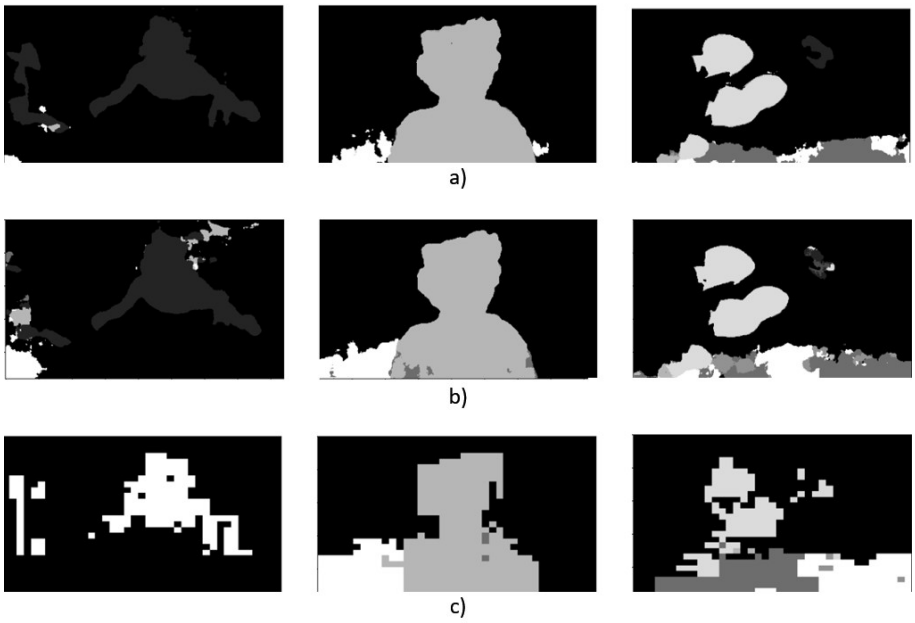


Figure 67: Segmented image after 30 evaluations by SegNet (a), PSPNET (b) and FCNN_16s.

It can be seen that PSPNET and SegNet already show the correct shape of the objects, albeit with better results for SegNet. Otherwise, the result obtained with FCNN_16s is of poor quality. The ability of a model to learn during the training process is very important, and since SegNet performed best in the final result, you can see an example of the evolution of the results in Figure 68.

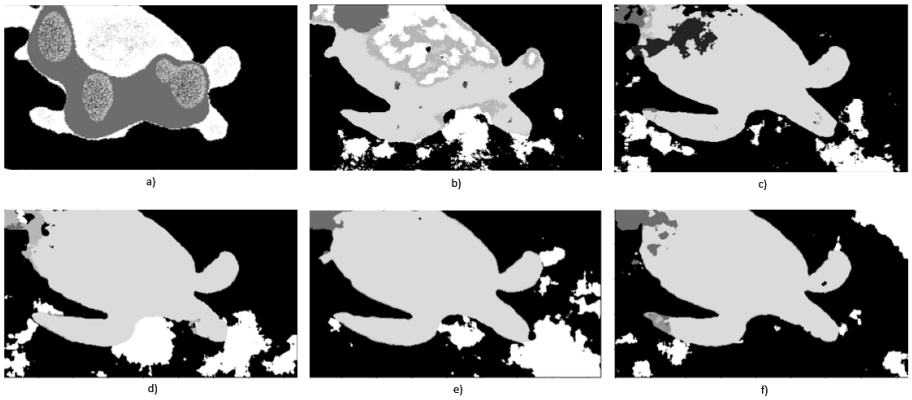


Figure 68: Evolution of the results in the learning process before: 200 training iterations (a), 1500 training iterations (b), 15000 training iterations (c), 45000 training iterations (d), 75000 training iterations (e) and 100000 training iterations (f).

The quality of the segmentation increases, but it is important to mention that after 15000 training iterations the result is very good, as the shape of the turtle is visible and the majority of pixels represent a well identified class. Otherwise, you can see in the Figure 69 the result of a bad segmentation before the 100000 training iterations, also

with the SegNet model. It can happen that there are 5 classes in the image and the result is confusing. The shape of the ship is unclear and the small classes, such as the fish on the left, are not visible in the result, which, in addition to the poor shape of the objects, also indicates errors in the correct classification of the individual classes.

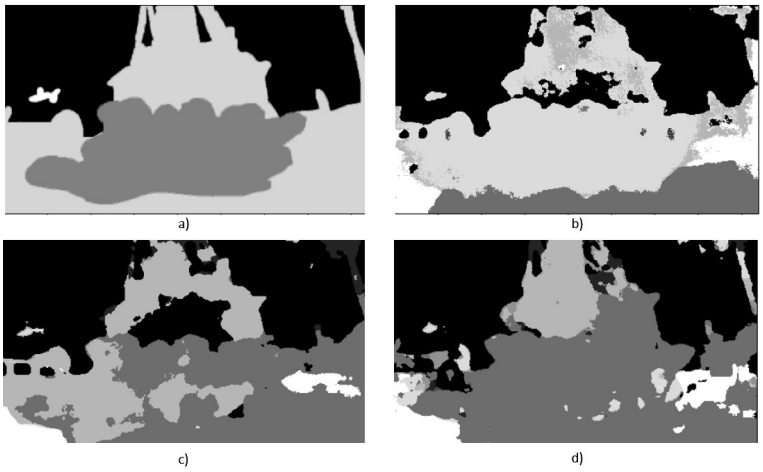


Figure 69: Evolution of the results in the learning process of the original image (a) before 15000 training iterations (b), before before 45000 training iterations (c) and before 100000 training iterations (d).

If we analyse the mean value of accuracy, we can see that the value decreases, which means that there are some classes whose accuracy is lower. This assumption is confirmed if we look at the column of the mean IoU value, because it shows that in addition to a poor prediction in relation to the predicted object class, a poorly segmented area of this object may also be possible. For this reason, it is necessary to evaluate the performance of the classification of each class by analysing the IoU metric, as it is the better metric to reflect the performance in the segmentation processes. Table 16 shows the results obtained for each class in relation to this metric.

Table 16: IoU [%] per class for 66 evaluations after 100000 training iterations.

	IoU							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
FCNN_16s	80.3	43.7	19.2	58.0	5.5	49.9	41.7	59.7
FCNN_32s	80.2	41.8	9.5	58.5	5.4	49.7	42.0	58.3
SegNet	86.1	54.3	28.7	60.2	8.6	49.6	55.8	61.5
PSPNET	87.9	56.0	13.5	54.0	7.2	42.6	45.8	63.4

If we look at the value of this metric in each class, we see that in most cases the best results are obtained for PSPNET and SegNet. The classes that are the most difficult to learn with the method are classes 2 and 4, as some require more training to achieve better results. These

poor results in both classes can be explained by the number of image examples available for the training process, which are far too few compared to the demand. Furthermore, in classes such as robots and instruments, the number of pixels of the object can be low compared to other classes such as wrecks. This fact can also be seen, for example, in the fish class, where each element of this class can be quite small compared to the size of the image. Otherwise, there are some classes that are the easiest for the algorithm to learn (classes 0, 5 and 7), as the results of a training procedure with validation after 200 training images each show. In the first validation of this experiment, the results for the first class are 68% of IoU and for the other two classes more than 30% of IoU with the SegNet model. As explained above, SegNet has the best ratio between performance (overall accuracy and mean IoU) and time spent, achieving better results than PSPNET in 5 classes. For this reason, all the experimental tests presented above were performed only for this method, with the main objective of improving the results and testing different parameterisations and critical features.

Time Optimization

In the initial optimisation, an attempt was made to reduce the time of the training process, because although training is an offline process, the time required should be as short as possible to ensure that it is not very difficult to add new classes to the process and that it does not take longer to improve the knowledge of an existing class. The first experiments are run with 0 workers for all models. This configuration means that the retrieval of the data is done in the same process in which the data loader is initialised, so it is not necessary to run multiple processes. If you test the number of workers with a positive integer, the data is loaded in several processes with the specified number of loader worker processes. Since the size of the dataset is not small, the change in the number of worker processes was tested, see results in Table 17.

Table 17: Accuracy [%], IoU [%] and time results when varying the number of workers.

	Accuracy		IoU	
	Overall	**Mean**	**Mean**	**Time**
0	77.4	49.8	39.3	56m36s
2	77.5	51.9	40.8	43m18s
3	78.1	51.6	41.0	43m45s
6	78.9	52.4	42.1	44m16s

As expected, using more worker processes speeds up the process a little, but using 2 or 6 worker processes is similar in time (differences

of less than a minute in total), but it is chosen to work with 6 worker processes because there is a small performance gain. The second attempt to reduce the time required for the training process was to change the size of the images. The first experiments were carried out with a resolution of 420 × 240 pixels, because it is known that for underwater applications the processing of the images has to be done with higher resolutions. Four different image sizes were therefore tested for 10 analyses, each with 1500 training images, and the results can be seen in Table 18.

Table 18: Accuracy [%], IoU [%] and time results varying the size of the training images.

	Accuracy		IoU	
	Overall	**Mean**	**Mean**	**Time**
120 × 60	67.6	42.4	31.3	23m22s
240 × 120	73.3	48.5	37.3	31m45s
420 × 240	77.4	54.8	42.7	1h10m35s
525 × 300	74.9	44.2	33.9	1h53m38s

From this it can be concluded that the best size of the images has now been used up, as their current overall and average accuracy is at its best. The lower the resolution of the images during training, the lower the performance, which is easy to explain as lower resolution images have fewer pixels, resulting in less information and lower resolution images. With large images, performance is lower because the images are stretched and the pixels can present confusing information. It is important to point out that this image size (525 × 300 pixels) in the IoU gives better results in the classes that are difficult to segment because they are classes whose size is smaller compared to the others. The size of the images is therefore inevitably related to the time required. For the reasons mentioned above, the remaining experiments were carried out with an image size of 420 × 240 pixels.

Performance otimization

In terms of optimising performance, it is necessary to check whether the configurations used for the training process are the best and whether it is possible to improve them. Regarding the optimiser, the repository can use some of the existing ones, but in the literature the optimiser Adam shows good results when it is well configured. It implements a stochastic gradient descent method based on an adaptive estimation of the first and second order moments. According to the literature, it is computationally efficient and requires little memory. It is suitable for problems with a large amount of data and/or parameters and its hyperparameters are intuitive to interpret and usually

require little adjustment. Nevertheless, it is important to consider which values are most appropriate for these parameters.

There are two crucial parameters for this optimisation:

- The experiments started with the default value for the **Learning Rate** of e^{-4} and then different values were tried with the main goal of finding the optimal learning rate. The experimental test was performed over 13500 training iterations, i.e. 9 evaluations, and the results obtained can be seen in Figure 70.



Figure 70: Comparative performance results at different learning rate.

- The **Weight Decay** is a common regularisation method for neural networks that helps neural networks to learn with good generalisations. Different values are experimented with during 13500 iterations, see Figure 71.

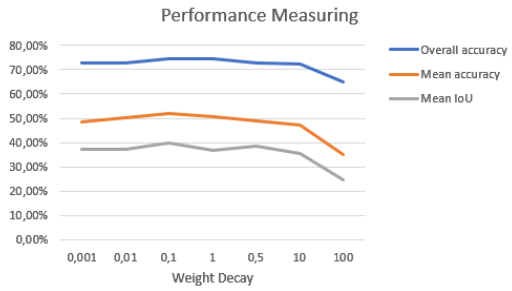


Figure 71: Comparative performance results using different weight decay values.

Since the metrics are consistent, i.e. if one metric goes up, that means three metrics go up, you can draw the conclusion by simply looking at these graphs of the last of 9 assessments. As you can see, there is a peak in the learning rate at e^{-4} and after that the performance drops abruptly. In terms of weight decay, the results are not very clear, but it can be observed that the performance drops a bit from 0.1 and the results are worse compared to this value. Therefore, this value was used for the next experiments. After some optimisation of the parameters, it was necessary to check the performance accuracy for a

long training process. Another important parameter is the number of epochs, i.e. the number of runs of the learning algorithm through the entire training dataset. An epoch consists of one or more batches, i.e. evaluations. Few epochs can lead to an under-trained model, while too many epochs can lead to over-fitting, where the model becomes too specialised on the training data. Therefore, 400 evaluations were performed after every 500 images. The results of this evaluation are shown in Figure 72. As you can see, the better result in this case can reach 80.82% in the maximum of the overall accuracy and outperforms the 50% by reaching 53.02% in the best result in the mean IoU, which is a better result than the first experiments. The main difference between this result and the first one, apart from the fact that the number of training iterations was doubled, is that the evaluation is performed every 500 instead of 1500 trainings. In the intermediate tests it could be seen that this configuration outperforms the result for some classes.

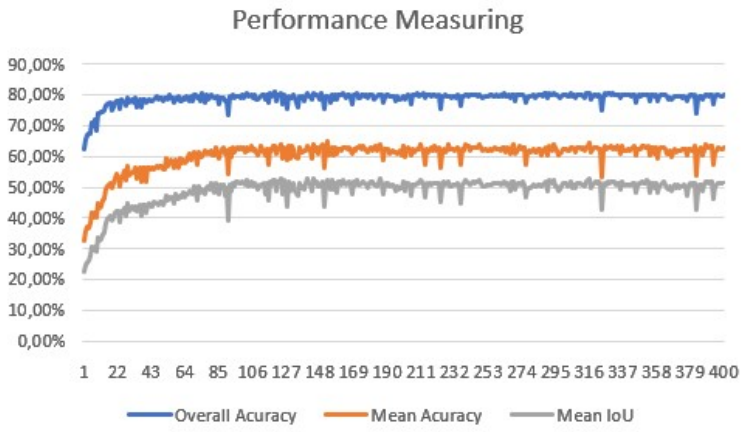


Figure 72: Performance evaluation over 200000 training iterations, with evaluations after every 500 iterations.

Table 19 compares the results achieved in the iteration number 100000 in both cases: Evaluation after 500 training images each (A) and after 1500 training images each (B). From this it can be concluded that in most classes the value of the mean IoU is higher and it is advantageous to evaluate and learn the model for all 500 images.

Table 19: IoU [%] per class after 100000 training iterations, if the evaluation is carried out after every 500 training images (A) and if the evaluation is carried out after every 1500 training images (B).

	IoU							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A	86.1	54.3	28.7	60.2	8.6	49.6	55.8	61.5
B	86.9	59.6	28.8	57.7	8.2	54.4	56.6	59.1

The Figure 72 shows that the values of the performance indicators are constant from fifty evaluations onwards, with small fluctuations. In this case, in terms of the overall result, it is not obvious that more training leads to better results, which could mean that the model cannot learn the classes in their entirety. This could be due to the fact that each class has a different number of images in the training set and the images may have different conditions, as well as the difficulty level of the test images (as shown previously). It is important to note that the mean value of IoU is usually smaller than other performance measures, which is due to the difficulty of correctly assigning all pixels of each class in each image. When we compare the results of 100000 and 200000 training iterations, we see that classes such as human divers, fish and rocks outperform the mean IoU results with more training iterations. If you look at the test images, their difficulty becomes clear. For this reason, an experiment was conducted with some images that were simpler: less noise and more contrast. The feature that was kept was the fact that the images contained different classes, because in practise there are often more than 2 classes of objects in the image. Another change was the number of images with the Plant and Robot classes, which make up more than 10% of the total images in this test set, instead of 33% because they are the less representative classes in the training set. In a test with 45000 training iterations and 90 evaluations, the value of the performance metrics is higher than that previously achieved in experiments. The overall accuracy increases by 83% and the mean IoU value by 54%. For some classes, the IoU value increases sharply, e.g. for classes 0, 3, 5 and 7, while the other classes show the same values as in the large training process. This proves that the difficulty of the test set can sometimes deviate too much from the intended conditions.

Supress the worst class

If we look at the results of each class along the different experiments, we notice that the worst classes to learn in this group of classes are undoubtedly the robots and the plants. As mentioned in the presentation of the dataset, the unbalanced number of samples per class can cause different types of errors and difficulties in the learning process. Otherwise, it is crucial to note that in the 110 images used to test the learning process, more than one third of the images contained these classes, while in the training images only less than 23% of the images contained these classes. Since the Robots and Instruments class is one of the classes that is less needed for the intended context, i.e. training a robot to recognise classes that may be present in the harbour area, it was decided to test the same dataset but to change the elements of the Robots class to the Human Divers class, as these instruments are often together with or close to the human divers. For this purpose, the

ground truth of the segmented images was changed, see Figure 73, and a training procedure was performed with evaluations in the original test set after every 500 training images.

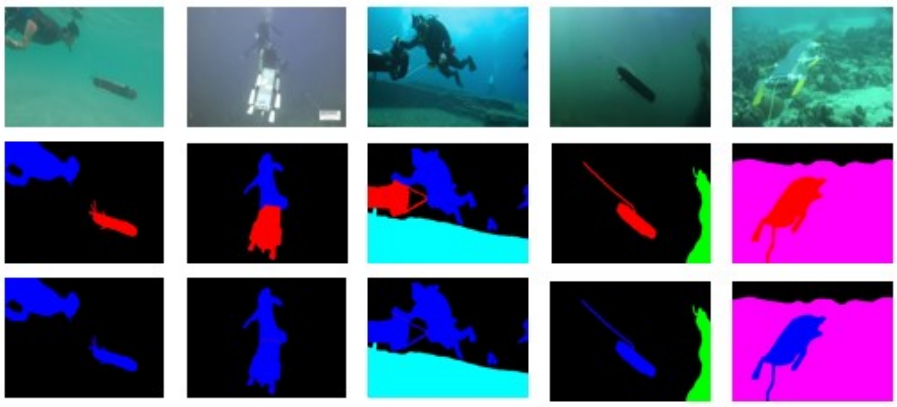


Figure 73: Illustrative images of the dataset without robot class.

The results after 100000 iterations of the training and 200 evaluations are shown in Table 20.

Table 20: Performance results [%] after 100000 training iterations without robot class.

Accuracy		IoU							
Overall	Mean	Mean	0	1	2	3	4	5	6
81.3	64.9	54.8	88.4	67.4	59.3	4.1	45.9	55.5	63.2

As you can see, the performance values generally increase when the robot class is deleted. The overall accuracy increases by over 3% and the mean IoU by over 5%, which is a good result in real-world contexts and applications. Looking at the individual classes, it can be seen that the human diver class greatly increases the value of IoU, as a combination of both classes (1 and 2) is present in the training images and the class is easier to learn in this configuration. The other classes show more or less the same result, with an increase in the wrecks class and the seafloor class. This could be related to the fact that these are classes in which the number of pixels is higher and that a model learns faster and better when the number of classes is reduced.

6.3.3 Discussion

The main objective of this work was to evaluate visual segmentation methods that have already proven their good performance in the literature, but can only be tested under outdoor conditions. The use of an autonomous vehicle in the underwater environment often requires the robot to know and recognise some of the objects that might appear.

Image segmentation allows not only to recognise the objects but also to extract some important information about them if this is intended, such as the shape, some assumptions about the dimensions, etc. The SUIM dataset used proves to be complete to perform the comparison, even if it presents some problems common to this type of datasets. Therefore, a series of experiments were planned and performed under the same conditions to find out which of the methods is best suited for the intended context. Although the results are not perfect, it is worth mentioning that all methods give good results for this real context, but when all measures and visual results are analysed and taken into account, SegNet stands out. This method has a higher capacity in the learning process and was therefore selected for the next tests. After some optimisations were made, a higher training was performed, but the results are similar. However, it can be seen that the results are more favourable when the test was performed after every 500 instead of 1500 images. Thus, the final results reach an overall accuracy of 80% and the mean value of IoU reaches 53%, even with very difficult test images. If we look at the results of the classes, we can see that there are two classes that are difficult for the model to learn and that the results are poor (less than 30% for robots and less than 10% for plants). This fact can be explained by the unbalanced dataset, which contains only a few images of these classes with a low diversity of perspectives and size representations. Nowadays, semantic segmentation is increasingly used in robots for underwater exploration, especially for autonomous navigation when the robot needs to recognise the elements it encounters during the mission, for example to avoid collisions. Therefore, it will be necessary to test some data strategies to see if it is possible to improve the result of the different classes, especially for the worst ones (plants and robots), which could be necessary in the intended context, because although the results presented in real contexts are not bad, it could be dangerous to work with them in a real application.

6.4 IMPROVING SEMANTIC SEGMENTATION PERFORMANCE

Although the results obtained in real contexts and presented so far are not bad, it could be dangerous to work with them in a real application and the question arises: *"Can we achieve higher accuracy in semantic segmentation?"*. According to the literature, it is possible to improve the results of semantic segmentation by adjusting the parameters of the model, resorting to some image enhancement techniques, increasing the number of images for training, using more balanced datasets, etc.

The first goal was to find the right parameters, and this was achieved in the first iteration. Now the next strategies need to be tested to see if they are successful in the underwater environment. So the main purpose of this work was to find out if there are some approaches to increase the quality and diversity of the training data commonly used in outdoor environments to improve the results of segmentation in the underwater context (see Figure 74).

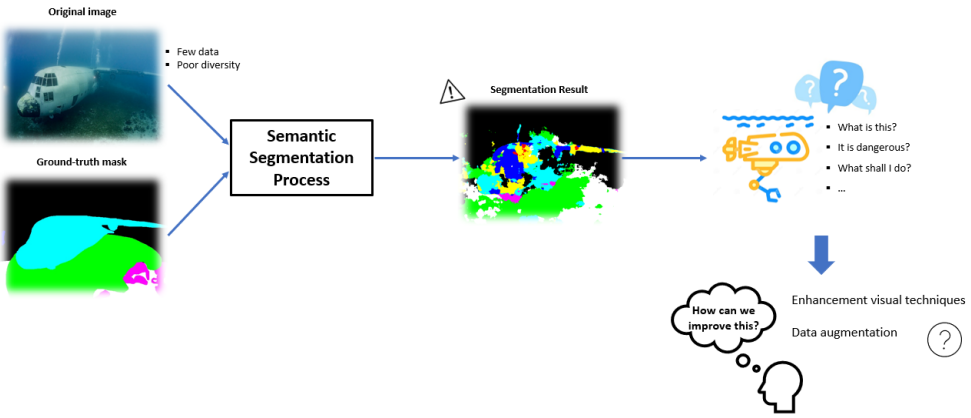


Figure 74: Diagram of the general framework of the proposal for improve the semantic segmentation results.

The main objectives of this study are therefore to evaluate the impact of the enhancement methods on the semantic segmentation results and to explore data augmentation strategies that allow to increase the number of training images and thus the diversity of the original dataset.

6.4.1 Methodology

As already mentioned, Segnet emerged from a series of experiments conducted under different conditions to find out which method of object segmentation is best suited to the context in question. However, it is now necessary to check whether the performance achieved actually delivers the best possible results. Therefore, a new experiment is conducted to find a starting point for the performance. It is important to note that the test dataset is difficult as it contains the main challenges encountered in underwater visual data collection, such as: inconsistencies in lighting, boundaries of objects that are not perceived as a whole, etc. The test is conducted after every 500 randomly selected images and a total of 100 assessments are made to obtain the final model and performance metrics. However, the best model was found after 45000 images, i.e. after 90 evaluations. Table 21 summarises the results in terms of overall accuracy, mean accuracy, mean IoU and the values for this metric for each class.

Table 21: Performance measurements [%] for the best result obtained with the original dataset in a total of 100 evaluations.

Accuracy		IoU								
Overall	Mean	Mean	0	1	2	3	4	5	6	7
79.7	64.4	53.1	87.2	62.7	29.5	58.8	16.6	51.6	59.1	69.6

As you can see from the table, although the results are generally not bad and useful, there are errors in some classes that are dangerous from the point of view of the autonomous vehicle. For example, the Plant and Robot classes have IoU values of less than 30%, which can lead to some erroneous actions. In the initial experiments, it was found that this can happen because these classes have a lower number of images for training compared to the other classes. Figure 75 shows the results of semantic segmentation for 5 examples. In general, it was found that the boundaries of the elements were correctly selected, but some of them were assigned to the wrong class, such as the diver in the second image or even the wreck in the fourth image. Although the results obtained are not bad, it is important to investigate if there is a way to increase these numbers to avoid possible errors. Since the study of the model's parameters has already been completed, the next steps will be to investigate other strategies that can be applied to the training dataset.

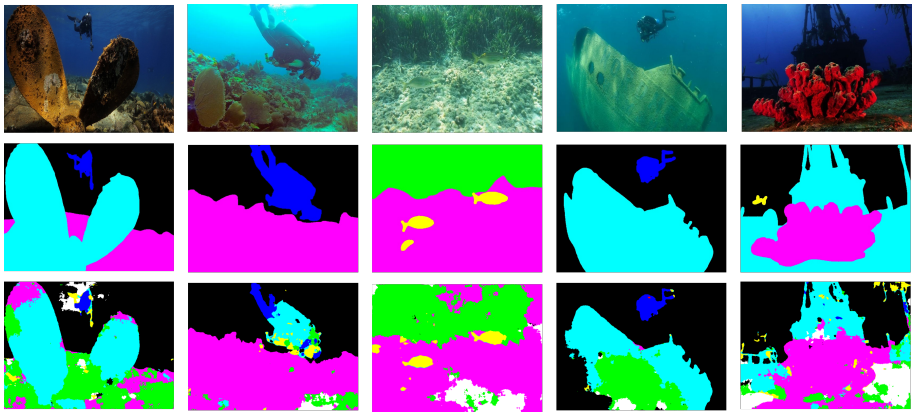


Figure 75: The segmented images obtained with the original dataset for a total of 100 evaluations after 500 training images each, based the best model.

As mentioned above, visual data collection in this world is not easy, as it is an expensive task that requires appropriate materials and communication platforms and is often dangerous for divers. Moreover, this task brings several problems, such as light absorption, turbidity and limited range, which reduces the number of good quality images. As already mentioned, the dataset in this particular case contains fewer images per class. Furthermore, this dataset is very unbalanced, as

the class with the highest and lowest representativeness is present in more than 80% and less than 10% of the images, respectively. Some of the images have worse visual conditions, with poor boundaries and without correct information about the colour of the objects. In other cases, there are some inconsistencies in the labelling mask, as this is a difficult and repetitive task. In Figure 76 you can see some of the problems of the dataset SUIM.

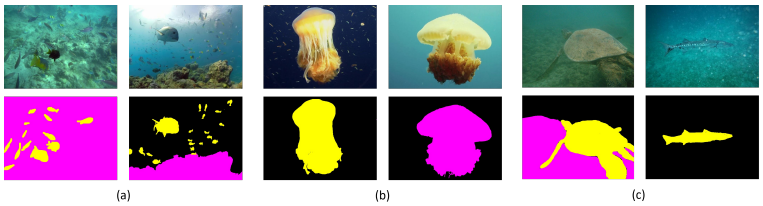


Figure 76: Problems of the SUIM dataset: more elements of a class in the original image than in the mask (a), some labeled errors when referring to the same class (b) and poor quality of the images (c).

Two examples can be observed where there are more fish in the original images than in the ground truth masks (Figure 76 (a)), which can be explained by the fact that labelling is a difficult task and is often done manually. Another example is the same class with different labelling in the masks (Figure 76 (b)), which can confuse the learning process. Finally, the third example shows the difficulties that often occur when perceiving the boundaries of the objects (Figure 76 (c)). Therefore, it is necessary to review the weak points of the dataset and improve some problems and solve others to improve the final performance (e.g. the poor segmentation).

According to the literature and everything that is known about the visual conditions of this environment, there are high expectations of the improvements that the enhancement of visual data will bring to the results of all processes that need to manipulate underwater images, namely semantic segmentations. According to the current state of the art, there are several techniques for different enhancements. One of them deals with the restoration of true colours, another with filtering techniques to denoise the image, and others that allow to remove backscatter caused by tiny particles in the water that reflect light back to the camera [90]. After some research in the literature and existing knowledge, it was decided to test only a few techniques, most of which enhance contrast and emphasise the boundaries of objects (crucial for a good training process) and which have already worked well in another related task such as feature extraction. Therefore, three techniques were tested that seemed suitable for the intended application: Contrast Limited Adaptive Equalisation (CLAHE) [91], White Balance (WB) [92] and a third technique [93] derived from diving experience that summarises some enhancements suitable for underwater images.

Thus, in detailed way:

1. CLAHE is used to increase the contrast of an image by redistributing the light across the image. It is one of the methods of histogram equalisation, but is not applied to the entire image, but only to small areas to reduce noise and provide a method that combines the different tiles without their boundaries being visible. This method is very important in underwater scenarios, as there are different lighting variations and in some cases a lot of darkness, and it promises to improve the visibility of object boundaries and some important features. The implementation used is based on the library Open Computer Vision Library (OpenCV) and requires some parameterisation. After some experimentation, a higher image quality was found by dividing the images into tiles of 80×80 pixels and using a clip limit of 1.

2. WB is a process that adjusts all pixel colour information so that it looks natural to humans. This is not an easy task because water absorbs and scatters light differently than air and many colours disappear in the depths where everything looks blue or green. In this context, this technique is used as post-processing, but it is not trivial to achieve the right white balance. There are several variants for implementing this technique, and the approach used shows that realistic colour degradation can be accurately minimised (without creating unrealistic colours, such as the purple of the grass, which can affect the results of semantic segmentation). However, the implementation used was not tested in underwater environments, which is why the experiments produced some images without quality. Different parameter combinations were therefore used and when a pattern of behaviour emerged in the results, the better combination was chosen.

3. The third method compared in this study, referred to here as dive correction, is a tool that can be used in code, as a desktop tool or in a browser. It consists of an image editor that adjusts each colour channel to restore the true colours. Basically, it is a simple process that goes through a few crucial steps: First, the average colour of the image is calculated, then a hue shift is applied to the red channel (up to an average red value of at least 60). After the RGB histogram is created with the new red colour, it is normalised using the lower and upper thresholds and then the colour matrix is created based on the new values. To summarise, this process makes the image appear more natural and contrasty.

It is also important to mention that since these methods only change the aspect of the image, it is not necessary to make a change in terms

of ground truth, and for this reason the original masks were used. As can be seen in Figure 77, the results of the enhancement methods vary in comparison to each other and it is necessary to check whether this pre-processing improves the final result.



Figure 77: Comparison of the improvement methods using 6 original images (a) with the CLAHE method (b), white balance transformation (c) and dive correction (d).

Looking at all images of a dataset that have been enhanced is a difficult and very subjective task. Therefore, Image Quality Assessment (IQA) methods have been developed over time to quantify the perceived image quality [94, 95]. Since image quality is a very subjective topic, it is important to teach the algorithm to evaluate good and bad qualities. These methods can therefore be used in an underwater context, e.g. to automatically recognise which enhancement method is the best. They can be categorised into 3 types: Full-Reference, Reduced-Reference and Blind or No-Reference. The first category requires a reference image (the ideal image in cases where the goal is to measure the increase in resolution of low resolution images) and is therefore based on the comparison of two images. In the reduced reference, there is no reference image, but some selective information about the image to be compared. In this underwater context, there are no reference images and for this reason these methods are discarded to draw conclusions. Therefore, there are some traditional no-reference methods that do not require a reference image and only the image to evaluate the quality. In this case, the idea is to simulate human visual perception and make quality assessments, such as:

- Blind/No-Reference Image QUality Evaluator (BRISQUE) [96], which is a spatial quality evaluator that finds and evaluates distortions. It deals with local contrast and luminance and is one of the most commonly used IQA methods. However, it does not

calculate distortion-specific features such as ringing, blurring or blocking.

- Perceptual-based Image QUality Evaluator (PIQUE) [97], which combines a set of statistical, structural and perceptual features tuned to human perception. The authors show that it is fast enough to be used in real-time applications.
- Natural Image Quality Evaluator (NIQE) [98], which evaluates natural images based on structural properties of the images such as luminance, contrast and textures. Sharpness and noise are also taken into account. According to the authors, it has great potential for use in unconstrained environments.
- Another metric presented in [99] relies on three types of low-level statistical features in both the spatial and frequency domains to evaluate image quality. It therefore quantifies the artefacts of an image and proves to be effective and efficient in evaluating image quality according to human perception. In this work, it is referred to as *sr_metric*.

Human perception is very difficult to mimic, but it is necessary to test these metrics in order to draw conclusions and better evaluate the results of enhancement methods when used in an underwater context. Consequently, it is possible to observe whether they are useful in semantic segmentation, because the better the quality of the visual images, the better the segmentation results could be.

Another reason for the poor results of semantic segmentation could be that the dataset is often too small compared to the demand and to obtain labelled datasets is expensive and time-consuming. Therefore, it is a challenge to train the models with small datasets, and this dataset contains only about 1500 images. Another approach to improve the results is data augmentation, which consists of various techniques that change the appearance of the images, e.g. rotating, zooming, flipping, colour variations, etc. In this way, the dataset can be enlarged. The use of data augmentation offers some advantages for deep learning methods, such as improving generalisation, increasing the robustness of the model to different variations of the input, using smaller datasets for some required applications and avoiding overfitting as the training data is more diverse. It is also worth noting that in most cases this approach is only used for the training dataset, as is the case in this study. In underwater scenarios, this is particularly important because, as explained earlier, underwater images present many challenges and the model has to cope with all the changes and variations. Therefore, it was crucial in this context to take into account lighting changes, distortions and changes in perspective. In addition, the water causes movement of the camera or objects during the shot and often additional

noise occurs as the conditions for the shooting equipment are often not the best, etc. Figure 78 shows some of the possible changes that can be made to an image to use it in the training set. So in this case, it is possible to get more images with just one image to better train the class of human divers. However, some of these changes are not useful for all scenarios, as they change the sense of reality, e.g. when the rocks appear at the top of the image.

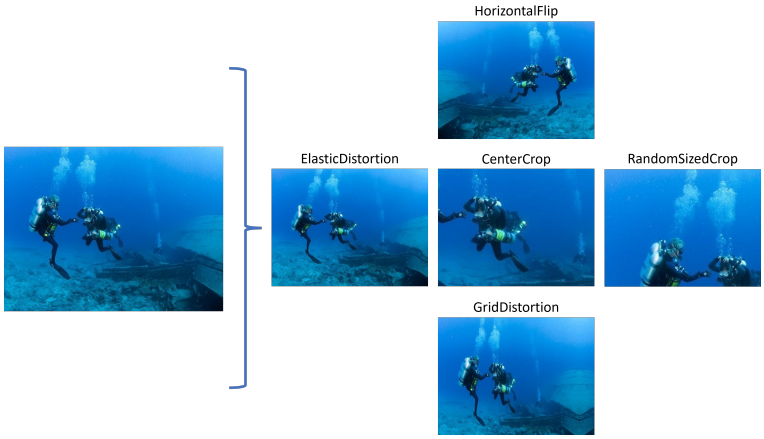


Figure 78: Five examples of possible changes to increase the size of the dataset using a single image.

As can be seen in Figure 79, vertical flipping is useful in some cases, especially when the camera takes images in an upper plane with respect to the object, i.e. images from the bottom of the underwater floor. But it is wrong if the image is taken in front of the object, as you can see from the air bubbles that can never come down. In many cases, extreme transformations (mirroring or rotating) must be avoided.

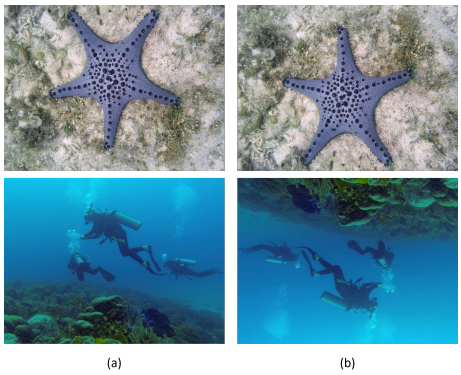


Figure 79: Result from two original images (a): in some cases, good variations result, but at the same time images arise that do not fit into the intended context (b).

For this reason, it is important to ensure that this process does not provide misleading information. The data augmentation in this application assumes that the images were taken from a vehicle that rotates

only slightly along the x axis. It is crucial that the consistency of the data provided is maintained. The changes must represent reality and the largest possible number of variations in the real world. There are already many methods of data augmentation to increase the size and variety of datasets. However, when a segmented dataset is required, the transformation must be applied to each image while keeping the ground truth consistent. Sometimes it is not trivial to maintain this in order not to add wrong information or e.g. black pixels during the rotation transformation. To solve this problem, it is often necessary to add a crop to the result of the first transformation. However, not all transformations require an intervention in the mask, such as changes to the colour information of the image where the position or size of the objects does not change. The various transformations can be divided into different categories, e.g:

- Spatial transformations - refer to changes in the image coordinate system that affect the position and orientation of objects and the overall image.
- Intensity transformations - change the appearance of objects, but not their shape, and act at pixel level.
- Elastic deformations - can simulate the flexibility of objects, the variability of their shapes or, in underwater scenarios, some water movements that are often seen in the image.
- Noise addition - makes the model more robust to the diversity of data collection, especially in underwater scenarios where there is noise in the images.
- Random erasing - some techniques remove regions and in the mask this class is ignored.

In the literature, there are several implementations of these transformations and libraries for the intended context, such as *ImageDataGenerator* of *TensorFlow* or the transformations of *PyTorch* when the user needs to apply the same transformation in the image and in the mask. However, these approaches have some limitations, such as the rotation of an image and the occurrence of black corners. Another option is Albumentations [100], a fast image augmentation library that applies the transformations to the images and ground truth simultaneously. It has an extensive set of transformations, from the most common to more specialised ones, such as grid-based distortions, and supports more than sixty transformations in total. It is designed to be adaptable and efficient when applying transformations in the training process. This ensures that both the image and the corresponding mask receive the same transformations and parameters. It also allows a pipeline of

transformations to be applied together, so the variety of transformations is incredibly large. For this reason, this library was used for this work. In Figure 80 you can see some transformations of the image and the ground truth that were performed by this library.

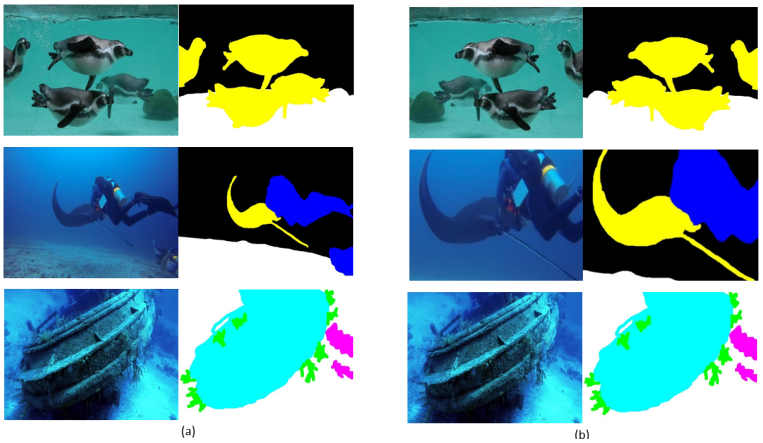


Figure 80: Some examples of original images (a) and corresponding transformations (b). The first and second rows represent transformations by flipping and cropping and the last row represents the deterioration in quality.

By flipping and cropping data, you can simulate different perspectives and sizes of objects that may appear on the images depending on the location and distance from which the image was taken. Therefore, the ground truth also changes in these cases. To better represent the different types of image acquisition systems, a downscale transformation can be used, as in this context some of these systems provide lower resolution images. In this way, the trained model is more robust in segmenting objects even in low quality images, but with this transformation it is not necessary to change the mask.

6.4.2 Experimental Results

In order to verify the statements made in the literature about the selected methods for increasing the final performance of semantic segmentation, it is necessary to test them individually and analyse their relevance to the intended context. In this way, new datasets are created for each test according to the respective assumptions. To compare the results, the common metrics and the quality of the visual results are used, as already explained.

Enhancement tecniques comparison

As already described, three implementations to improve the visual quality of the images are tested and compared: CLAHE, WB and the dive correction. It is expected that a better visual quality of the

images will lead to better training and consequently to better semantic segmentation results.

A correct evaluation of the quality of the obtained images is a difficult task, especially in cases where no image is available as a reference. However, regardless of the underwater conditions under which the images were taken, it is important to have an idea of whether these newly generated images lead to a more appealing result or not. As it is very difficult and subjective to check the result of all images in the dataset, some traditional metrics were used for blind quality assessment, i.e. without resorting to a reference image. Therefore, a set of 20 images of each enhancement to test with 4 previously selected metrics: BRISQUE, PIQUE, NIQE and *sr_metric*. An attempt was made to make the set as heterogeneous as possible in order to obtain a variety of contexts. This can be seen in Figure 81, where there are images with a quality score of more than 40 (corresponding to mediocre quality) and others with less than 20, i.e. excellent quality, according to the methods BRISQUE and PIQUE.

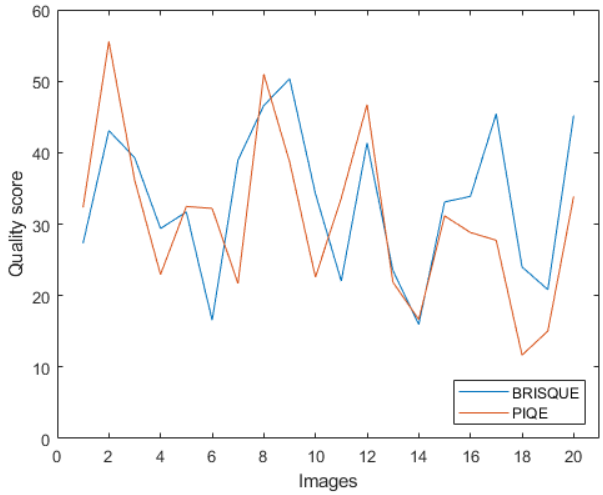
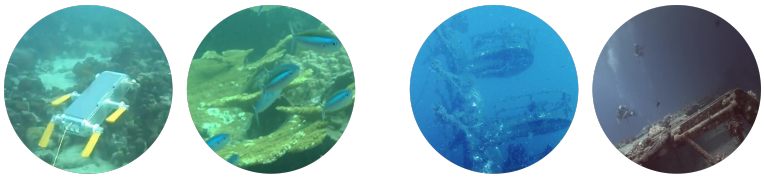


Figure 81: Quality assessment with BRISQUE and PIQUE for 20 original images selected to evaluate the quality of images processed with enhancement techniques.

The purpose of this type of test is to ensure that the most suitable method not only improves the images with low visual quality, but also does not degrade the already good images. If you look at the graph, you can quickly see that the results of these two metrics do not always agree, but both say, for example, that the 2^{nd} and 12^{th} images tested are of poor quality and that this is true in terms of human perception (Figure 82 (a)). Furthermore, both metrics indicate that the 14^{th} and 18^{th} images are of high quality, but this conclusion is not clear in terms of human perception, as you can see in Figure 82 (b).

To see if it is possible to determine which enhancement method is best suited in the underwater context according to these evalua-



(a) (b)

Figure 82: Two examples of bad visual images (a) and good visual images (b) of the original set, according with the methods BRISQUE and PIQUE.

tion methods, the quality tests are performed for the 3 enhancement methods and with the 4 selected metrics. The results obtained are shown in Table 22. The values obtained for the metrics that exceed the results obtained with the original set are highlighted in bold. It is also important to note that the BRISQUE and the PIQUE represent the results on a scale from 0 to 100, while the NIQE and the *sr_metric* are based on a scale from 0-10. In addition, for the *sr_metric*: the higher the score, the better the quality of the image.

Table 22: Comparative results of the mean of the quality score of the Dive, WB and CLAHE corrections for a total of 20 images, using PIQUE, BRISQUE, NIQE and the *sr_metric*. For each metric, the results that exceed the results obtained with the original set are highlighted in bold.

Data	PIQUE	BRISQUE	NIQE	*sr_metric*
Original	30.63	33.14	3.30	6.35
Dive	31.9	**30.16**	**3.15**	**7.27**
WB	36.7	37.45	3.38	5.92
CLAHE	32.50	**27.50**	**3.14**	**7.40**

For the other metrics, the relationship is reversed, i.e. the best results are achieved with a lower quality score. As you can see, for the first metric, the overall quality of the results obtained decreases as the quality score increases. For BRISQUE, the results of the dive and CLAHE corrections increase the quality of the images, with the CLAHE method being particularly emphasised as it greatly improves the average quality score (close to excellent quality, which is good for the underwater context). For NIQE and the *sr_metric*, these implementations also increase the quality of the images in general, but with a small increase for CLAHE. Looking at the individual results of the metrics obtained for the CLAHE enhancement method, it can be seen that for the 13^{th} image, the score for the quality of the obtained image increases sharply (more than 18 units)compared to the original visual quality for the BRISQUE metric. So this indicates a good result. However, when we analyse the results of the score for the same image

in the other metrics, the value of the quality score for PIQUE and NIQE shows a decrease of 6 and 0.3 respectively, indicating a deterioration in quality. The results obtained for the *sr_metric* do not differ from each other. Therefore, it cannot be said with absolute certainty that one method is better than the other for producing more appealing visual images. Although all results indicate that the CLAHE method and the dive correction provide the best results in terms of improving image quality, there are sometimes contradictions. Therefore, it was decided to test all obtained images with three enhancement methods to determine whether the method provides the best results for segmentation. The three methods are thus applied to the entire dataset of training images and also to the test images. For this test, four different datasets were created and the model was trained with a total of 50000 images, with the evaluations performed on 500 randomly selected images, resulting in 100 evaluations each. The results obtained are compared with the original dataset, i.e. with the training performed on images without the image enhancement already described, and can be seen in Table 23 and 24.

Table 23: Comparative results of best measured model performance [%] in terms of Accuracy (Overall and Mean) and IoU (Mean) achieved with: CLAHE, WB and dive corrections in a total of 100 validations, each after 500 training images. Whenever the results achieved exceed the results of the original set, they are emphasised in bold.

	Accuracy		IoU
Method	**Overall**	**Mean**	**Mean**
Original	79.7	64.4	53.1
CLAHE	80.4	64.0	**53.3**
WB	80.0	62.5	51.9
Dive	**80.5**	63.4	52.3

Table 24: Comparative results of best measured model performance [%] in terms of IoU for each class achieved with: CLAHE, WB and dive corrections in a total of 100 validations, each after 500 training images. Whenever the results achieved exceed the results of the original set, they are emphasised in bold.

	IoU							
Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Original	87.2	62.7	29.5	58.8	16.6	51.6	59.1	69.6
CLAHE	87.0	61.6	25.7	**59.1**	14.1	**54.2**	**60.5**	64.2
WB	86.6	58.2	**30.0**	**59.7**	8.5	**55.8**	56.8	59.4
Dive	87.1	58.2	27.4	**60.0**	11.6	**53.9**	57.6	62.8

In general, the mean IoU value is the best way to check the performance of the semantic segmentation methods, since it is calculated with respect to all classes. However, since we have 8 classes, it is normal that this value is not much higher. In other words, it indicates how well the segmentation model works across all classes. As for the quantitative results, it shows that CLAHE only slightly increases the mean IoU value. But it provides better results than the original for tree classes and in the case of the fish class with significant results (more than 3%). For this reason, this improvement method seems to be the best, but the visual results should be observed. Figure 83 shows the results obtained by applying the three enhancement methods to the images.

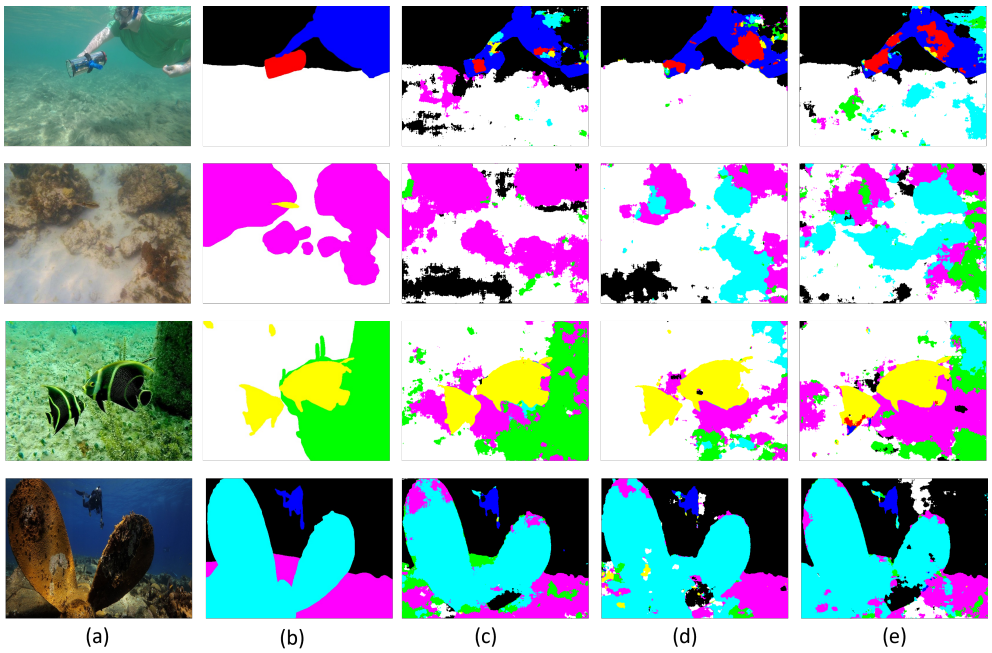


Figure 83: Qualitative comparison results of 4 different examples (a and b): the CLAHE method (c), WB method (d) and dive correction (e) using the best model in a total of 100 validations, each after 500 training images.

Looking at the visual results, it is possible that the results of CLAHE are generally better because it is the best method for recognising the human diver's body and the corals in the second example. Also, it is a unique method that shows the plants of the third example in a high number of pixels, which is a good result as it is a difficult class to model. The other implementation stands out compared to the CLAHE method in the robot class, but is not significant as it is only one of eight classes and does not provide good segmentation results. In this way, the best result is obtained with the CLAHE method, which slightly improves the final result and has a higher overall accuracy, i.e. a higher rate of well-recognised pixels. This result was to be expected as the contrast and therefore the visibility of the main features of the

images is improved, which can increase the performance especially for underwater images with different lighting conditions. It is also important to mention that this enhancement technique allows the model to reach the best value of mean IoU (within 50000) after 90 evaluations, i.e. using a total of 45000 training images instead of 50000 for white balance enhancement or 47500 for dive correction. For this reason, the CLAHE method is selected as the best and most suitable for the context and for use in the next experiments, if necessary.

Figure 84 shows an example of the improvement of the final segmentation results with the CLAHE method in the fish and wreck classes compared to the original results.

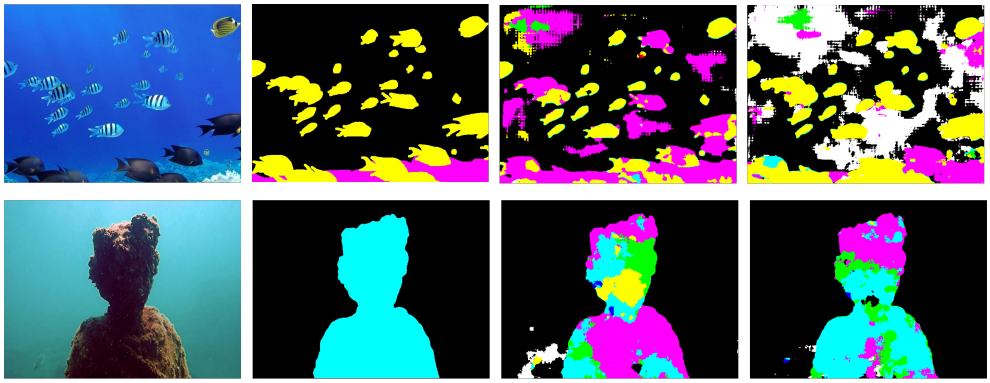


Figure 84: Comparison of the qualitative results of 2 different examples (a and b), between the result obtained with the original set (c) and the CLAHE method (d) using the best model for a total of 100 validations.

Data Augmentation Influence

As for Data Augmentation (DA) is concerned, there are several options that can be applied to the images, but it is not possible to apply all of them, as the dataset would then be larger than necessary and the processing time would increase significantly without necessity. To create the datasets with data augmentation required for the following tests, a total of eight transformations were used, which can be seen in Figure 85. Clipping operations were used as well as some distortions (grid and elastic) and a flipping in the horizontal axis. It is important to note that the test of the influence of data augmentation was performed in a isolated way, i.e. without taking into account visual enhancement methods. The first experiment with the Data Augmentation was performed with a dataset in which 5 of the augmentations were applied to each image of the training dataset. This resulted in a total number of 9144 images for training to evaluate the effects of this phase in a larger dataset, as 1500 images is a small dataset for the intended context.

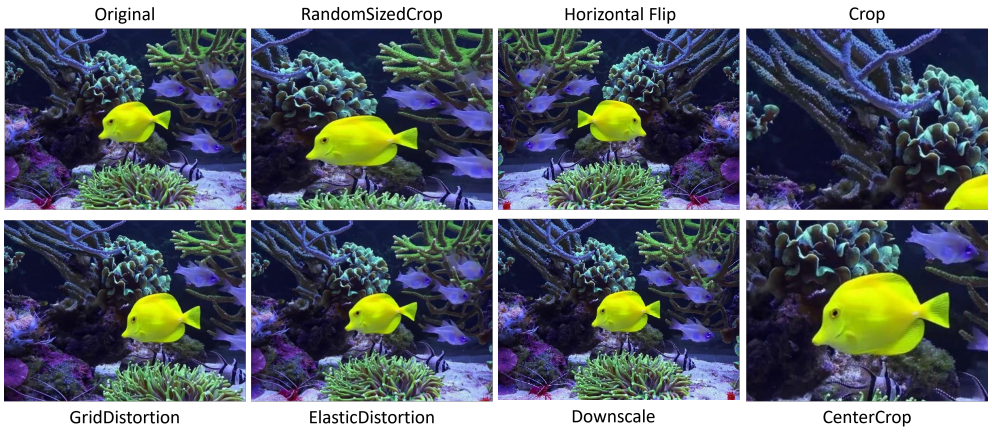


Figure 85: Eight transformations used in the creation of the augmented datasets for training.

Figure 86 illustrates the distribution of classes across the images. It can be seen that the number of samples per class increases and the distribution of classes remains the same.

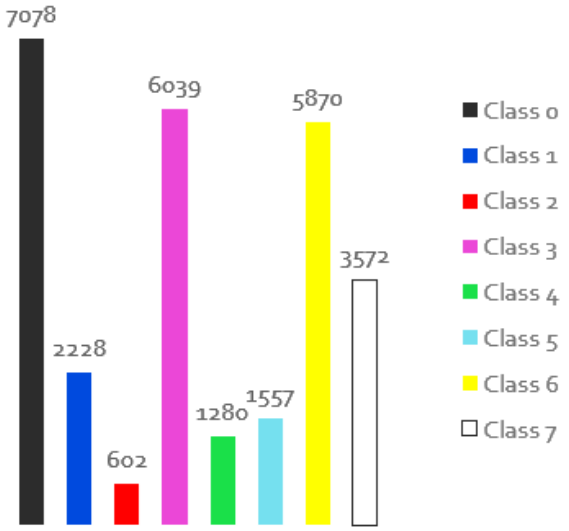


Figure 86: Distribution of the individual classes for the images of the training set with Data Augmentation.

To allow for a fair comparison and because the new dataset is larger, the original dataset was run for 50000 training images, with the analyses performed for 1000 images each (about two-thirds of the total number of images). Therefore, the new dataset was evaluated with DA after every 6000 images for a total number of 300000 iterations, with the aim of having the same number of evaluations at the end, i.e. 50 evaluations. The tables 25 and 26 summarise the result for the best iteration for the original dataset and the dataset with Data Augmentation (iteration 41 and iteration 47 respectively).

Table 25: Performance measures [%] for the best result obtained with the original dataset and the dataset with DA in a total of 50 validations. In bold type you can see that the results achieved with the data augmentation are better than the original results.

	Accuracy		IoU
Method	**Overall**	**Mean**	**Mean**
Original	80.6	62.9	51.8
DA	80.4	**65.0**	**54.1**

Table 26: IoU [%] for each class in the best result obtained with the original dataset and the dataset with DA in a total of 50 validations. In bold type you can see for each class whether the results achieved with the data augmentation are better than the original results.

	IoU							
Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Original	87.6	60.9	20.9	60.2	11.6	54.1	55.5	64.0
DA	85.7	**64.2**	**38.6**	**61.3**	8.6	**54.3**	**61.1**	58.8

If we look at the results, we see that increasing the number and variety of samples (under the same conditions for the train) increases the overall value of the mean IoU by about 2.5% (bold in the table), which is a relevant result since there are 5 classes whose IoU is increased. There are two classes where behavioural performance decreases, but one of them is a difficult class to model (plants) and another is a general class as it is rocks and the value of this class is still useful in the intended context. Figure 87 shows the results obtained during the learning process in terms of the mean IoU value. It can be seen that the result with data augmentation generally performs better throughout the process. If we look at the first 25 evaluations, we see that the difference between the maximum values achieved is about 6.5%, which means that Data Augmentation learns faster.

Figure 88 shows the same analysis, but in relation to the individual classes of the process. It can be seen that in most cases the data augmentation improves the result of the IoU values or starts better and faster than the results without Data Augmentation. With respect to classes 0 and 7, the DA results are not better than the original results, but these are common classes that are strongly represented on many images and therefore easy to learn. Regarding the worst classes (plants and robots), it is important to point out that the results start earlier with data augmentation, i.e. the ability to learn in the training dataset with a larger number of samples. These two values are lower than the other classes because the dataset remains unbalanced, meaning that these classes are less represented than the others.

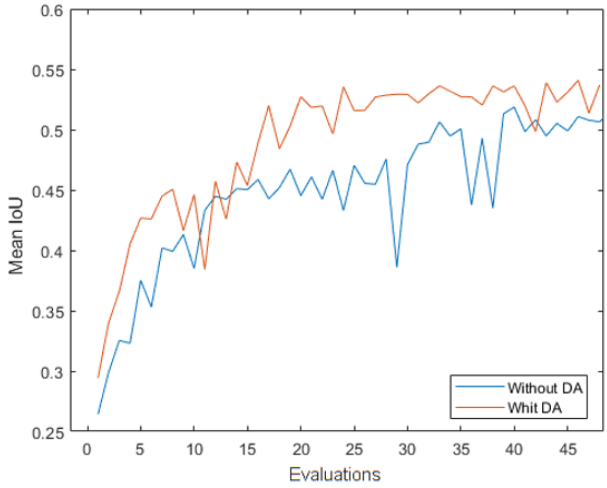


Figure 87: Mean IoU measurement along the 50 evaluations with the original dataset and the DA dataset.

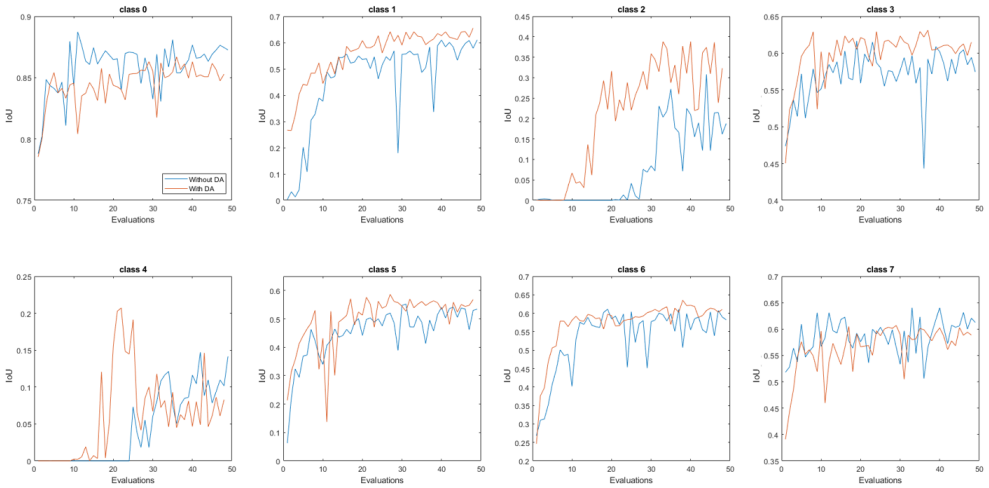


Figure 88: IoU measurement along the 50 evaluations with the original dataset and the DA dataset, for each class.

Figure 89 shows the results of the original dataset (second column) and the augmentation dataset (third column) with the same configurations. It can also be seen that for most classes and images, the result with the augmented data is better than the previous one. Pixels well identified as robots appear in large numbers, and wrecks and human divers are better delineated. Plants and robots are the classes that give the worst results, as they are still the classes with low representativeness.

It is clear that segmentation helps to improve the results as it generates more images in which the classes are represented with different perspectives, sizes and image qualities. So in the next experiment, let us see what happens when we apply the data augmentation to all images, but in a different number of variations depending on the class.

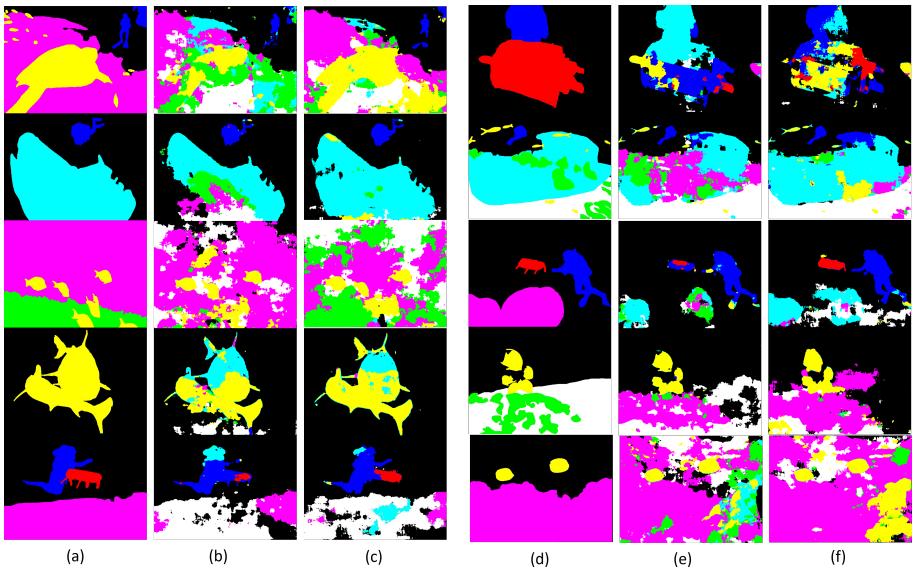


Figure 89: Qualitative comparison of the results of 10 different examples of semantic segmentation (a and d) using the original dataset (b and e) and DA set (c and f) for the training process, where the best model was selected in a total of 50 validations.

If we look at the distribution of the original dataset used for training (see Figure 65), we can see that the classes robots, plants and wrecks are less represented in the images. Therefore, two image data augmentations (RandomSizedCrop and HorizontalFlipping) were applied to all classes and another 5 transformations (CentreCrop, Crop on the left, Grid and Elastic Distortion and Dowscale) were applied to the less representative classes. The generated dataset comprises a total of 7312 images. This is a larger dataset, but it still needs to be checked whether this number of images is sufficient to model the 8 classes. The new distribution of classes in all images is shown in Figure 90.

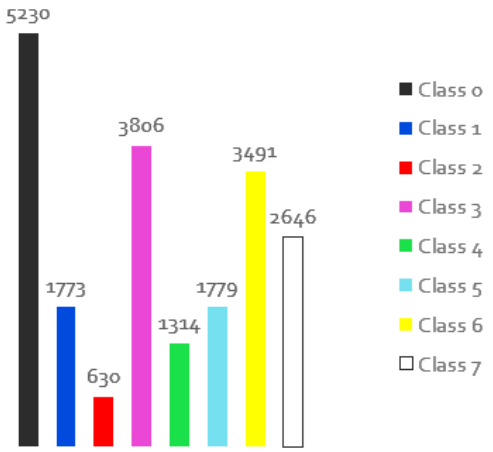


Figure 90: Distribution of the individual classes for the images of the training set when applying the data augmentation in different number according to the class.

It can be seen that the less representative classes are the same as in the previous experiment, but with a higher number of samples per class. The higher representative class is present in about 70% instead of 80% of the images and the three less representative classes are present on average in about 17% instead of 12%. Note, however, that the increase of the worst classes is not proportional (while the robot class increases by 6 times, the wrecks increase by 7 times) because the classes are not isolated in the image and some cropping transformations may suppress some of them. This factor, together with the different size of each class in each image, can influence this result. As in the previous experiment, training was performed on the images of this new dataset with a data augmentation (DA_v1) after every 5000 images (about two thirds of the total number of images in this dataset) for a total number of 250000 iterations, with the aim of ending up with the same number of evaluations as the original result presented previously.

The tables 27 and 28 summarise the result for the best iteration for the original dataset and for the dataset with the new data augmentation dataset during the segmentation process over the 50 evaluations.

Table 27: Performance measures [%] for the best result obtained with the original dataset and the new (DA_v1) dataset in a total of 50 validations. In bold type you can see that the results achieved with the data augmentation are better than the original results.

	Accuracy		IoU
Method	**Overall**	**Mean**	**Mean**
Original	80.6	62.9	51.8
DA_v1	80.2	**66.0**	**54.4**

Table 28: IoU [%] for each class in the best result obtained with the original dataset and the new with data augmentation (DA_v1) in a total of 50 validations. In bold type you can see for each class whether the results achieved with the data augmentation are better than the original results.

	IoU							
Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Original	87.6	60.9	20.9	60.2	11.6	54.1	55.5	64.0
DA_v1	86.3	**61.1**	**37.7**	**61.2**	**23.5**	53.4	54.8	57.9

As you can see, both the value of the mean IoU and the mean accuracy increase, as there are some classes that increase both the accuracy in recognising the pixels and the IoU. The classes that have been augmented in the training set increase the value obtained, with the exception of the Wreck, but the behaviour in this class is similar to

that of the original set. However, the Plants and the Robots increase the value of IoU by 12% and 16% respectively, which is good as they are difficult to model and learn. The class of divers is not greatly augmented compared to the robots, but they increase their IoU value as the robots appear in the image together with the human divers in most cases. The wrecks do not improve their results, although they belong to the augmented classes, but their results are already higher than 50% and for this reason it is a good result. A curiosity is that this model was chosen because it has the best value for the mean IoU during the 250000 training iterations, but this best value corresponds to iteration number 22 (fewer evaluations).

Figure 91 therefore shows the segmentation obtained for images with the classes added in the training process: Wrecks, Plants and Robots. For these specific cases, you can see that the results are much better than those obtained with the original set. In addition, the area of the plants is much better delineated and the bodies of the robots and the wrecks are well identified.

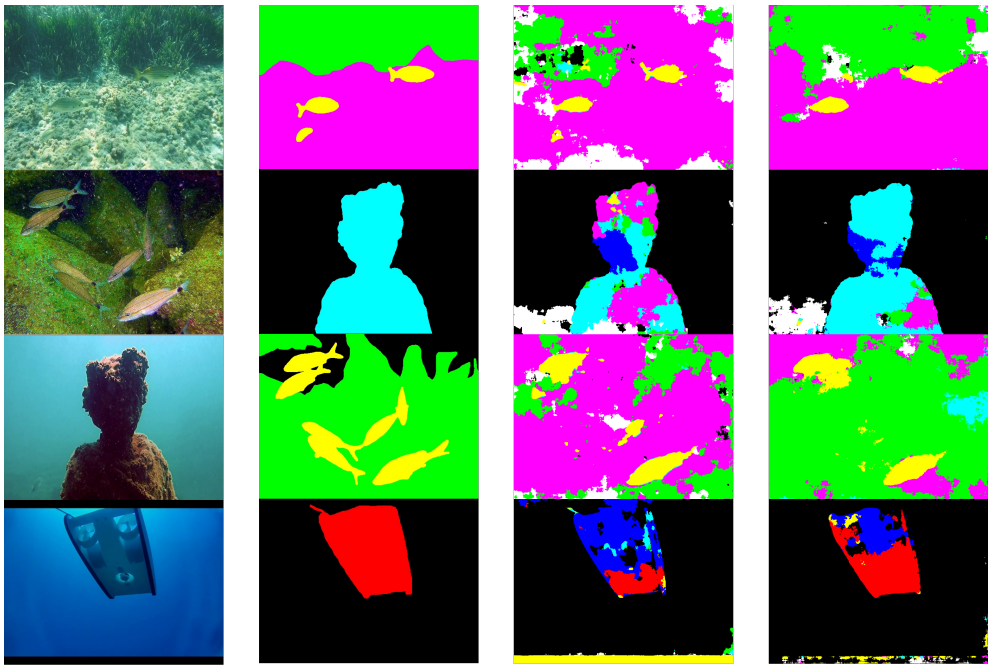


Figure 91: Qualitative comparison results of 4 different examples of semantic segmentation (a and b), obtained from the original training set (c) and the DA_v1 dataset (d), using the best model in a total of 50 validations for each case.

The last result, related to Data Augmentation, has the main goal of testing whether it is possible to increase the performance of an isolated class by data augmentation in the images when this class occurs. For this purpose, a dataset was created in which the images of the robot (class with low representativeness) were augmented with the 7 semantic segmentations of the previously described ones. The new dataset shows the distribution of the graph of Figure 92, so that a total of 2252 images were obtained. The number of images with the robot class increases sharply and their representation reaches more than 25%, which is a good result for a dataset with 8 classes to be segmented. The plant is the class with the lowest representation and the human divers increase their value to more than 30% as they frequently co-occur with the robot.

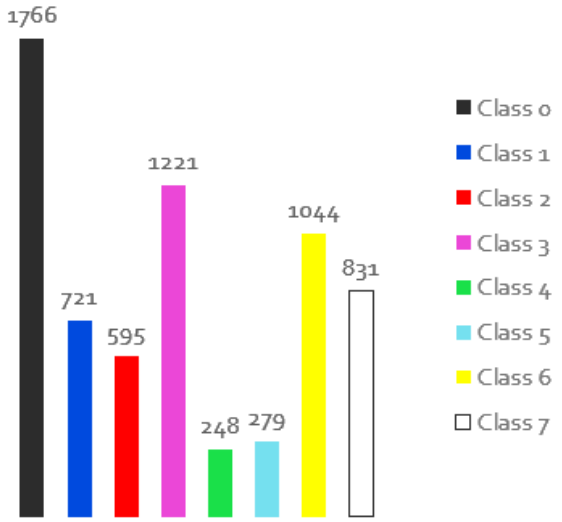


Figure 92: Distribution of the individual classes for the images of the training set when applying the data augmentation only in the robot class.

Therefore, this dataset was trained with a total of 75000 images, where all 1500 images were evaluated to perform a total of 50 evaluations, and the best model was determined after 36 evaluations. The result of this best evaluation gave an IoU value of 41.4% for the robot class, which is an incredible result for the class that had a maximum value of 20.9% in the original case. As expected, the value for the human diver also increases to 67.8%. These two increases lead to a higher mean value than in the original case (54.5% instead of 51.8%), proving that one or two classes with poor results can strongly influence the value of the mean IoU. Figure 93 shows the result for the value of the IoU of the robot class during the whole training process.

It can be observed that, in addition to the fact that the value in the original dataset is always lower than when using the images with data augmentation, the values in the dataset with data augmentation start to show up before. In the original case, the class only shows results that are higher than 0 after 25 evaluations (also unusable).

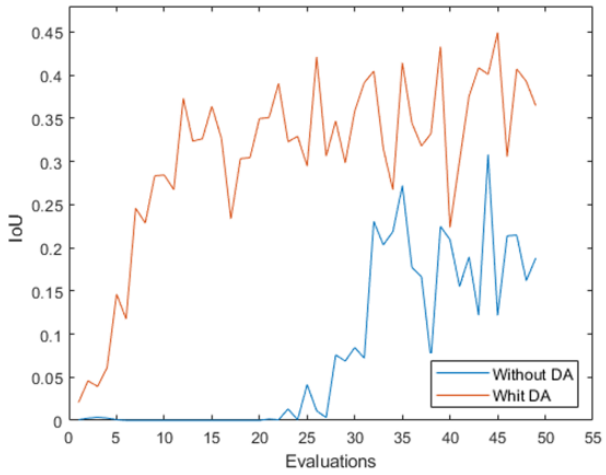


Figure 93: IoU measurement for the robot class (class 2) along the 50 evaluations with the original dataset and the Data Augmentation dataset.

Figure 94 shows the result of some images with an existing robot class for which the best model obtained in this test was used.



Figure 94: Qualitative results of different examples of segmented image with robots, obtained from the Data Augmentation set for the robot class, using the best model in a total of 50 validations.

The class of plants is also less representative in the original dataset. Therefore, a new dataset was created in which the images of the plant were supplemented with 7 data augmentations. The difference is that only about 100 of the total number of plants are supplemented, as this class occurs in many images with very small sizes or only in one corner of the image and the transformation generates images without this class (see Figure 95).

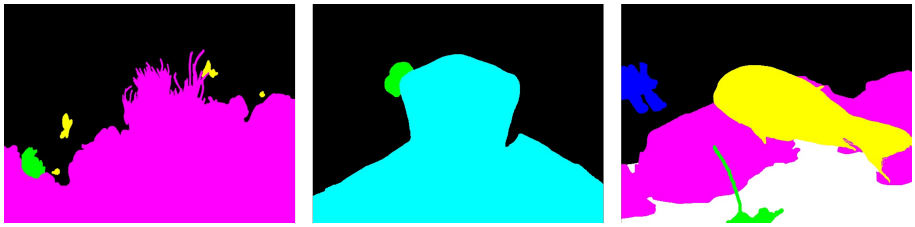


Figure 95: Demonstration of the size and location on the original set of the plants in some of the images.

The remaining images are retained in the dataset without data augmentation in a similar way to all images without a plant class. The result is a dataset with a total number of 2588 images, but with more images with the plant classes, this class being the fourth class most represented in all images, see Figure 96.

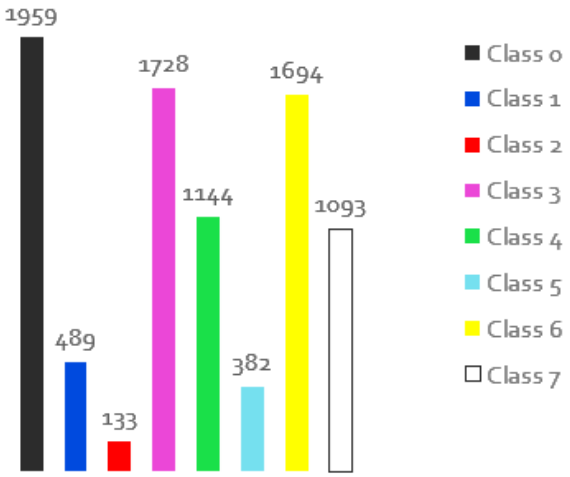


Figure 96: Distribution of the individual classes for the images of the training set when applying the Data Augmentation only in the plant class.

Therefore, this dataset was trained with a total of 75000 iterations, evaluating all 1500 images to perform a total of 50 evaluations, and the best model was obtained after 45 evaluations. As far as the result is concerned, a similar behaviour to the robot class can be observed, i.e. it starts with values above 0% and reaches higher results over the course of the 50 evaluations (see Figure 97).

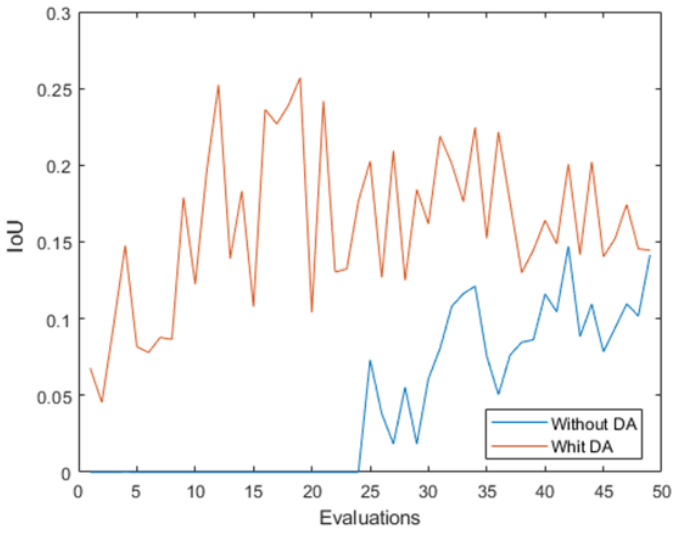


Figure 97: IoU measurement for the plant class (class 4) along the 50 evaluations with the original dataset and the data augmentation dataset.

In this case, the best value of the mean IoU (53%) does not refer to the best value of the IoU for the plants, since all classes are considered

in this metric. The value obtained for the IoU for the plant is around 20% for the best model, which is also a higher value than the results obtained without data augmentation. It is important to point out that this class is very difficult to model as it can appear on the image in different sizes, perspectives and variations.

Data augmentation with CLAHE method

To evaluate the effect of using both strategies simultaneously and to see if the results improve, it was decided to use the DA_v1 dataset, i.e. the more balanced dataset of the previous tests and the better enhancement technique (CLAHE). Therefore, training was performed with the images from this new dataset (DA_C)) after every 5000 images for a total number of 250000 iterations to obtain 50 evaluations. The Tables 29 and 30 show the comparison between the new results and the results obtained in the test with the Data Augmentation only.

Table 29: Performance metrics [%] comparison in terms of Accuracy (Overall and Mean) and Mean IoU for the best result with the more balanced dataset with data augmentation (DA_v1) and this dataset with CLAHE method (DA_C) applied for a total of 50 validations. In bold type you can see that the results achieved with the data augmentation are better than the original results.

	Accuracy		IoU
Method	**Overall**	**Mean**	**Mean**
DA_v1	80.2	66.0	54.4
DA_C	**81.8**	**67.6**	**56.0**

Table 30: IoU [%] for each class in the best result with the more balanced dataset with data augmentation (DA_v1) and this dataset with CLAHE method (DA_vC) applied for a total of 50 validations. In bold type you can see for each class whether the results achieved with the data augmentation are better than the original results.

	IoU							
Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
DA_v1	86.3	61.1	37.7	61.2	23.5	53.4	54.8	57.9
DA_C	**87.9**	**66.7**	**39.4**	**62.4**	12.8	**53.9**	**61.0**	**63.9**

Based on the results, we can see that all performance metrics are increasing, especially overall accuracy, which has been the best metric since the first test. Despite a lower score, the mean accuracy improves by 1.5% and the mean IoU reaches a value of 56%, which is the best result ever obtained with this dataset (a good result considering the number of assessments performed). If we look at the results of each

class, we can see that they are all increasing, with the exception of the value of Plants. It is important to mention that the value of IoU obtained during the training process for the Plants class has a maximum of 17.3%. This could be due to the fact that the CLAHE could damage the region of the specific plant areas on the images. Thus, it has been shown that both strategies together are a good plan to improve the results in general. Figure 98 shows the segmented image result for some of the images used to test the different approaches with the best model trained on this final dataset. The results show good segmentations and many of the pixels are well labelled (some of the classes, such as robots or fish, represent the complete object).

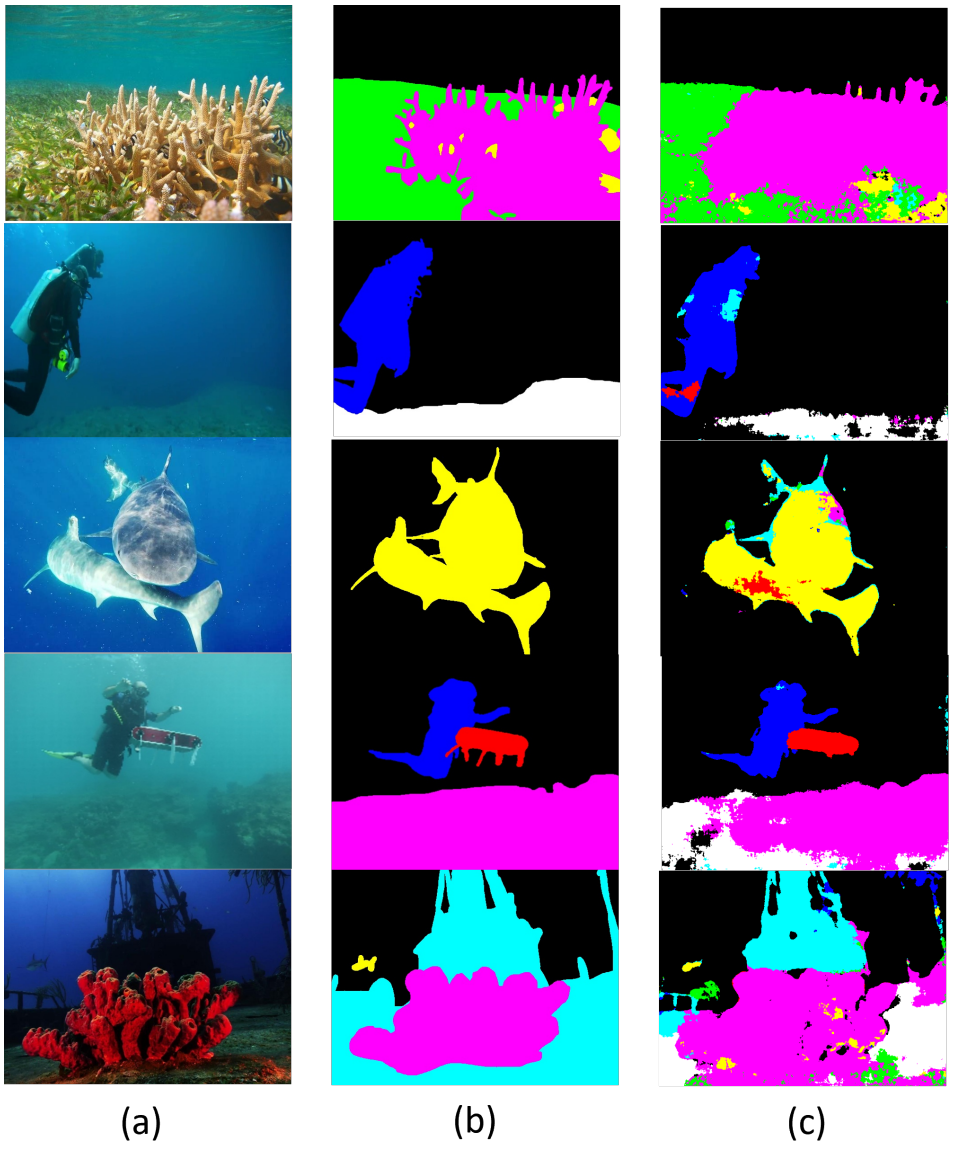


Figure 98: Results of 5 different examples of semantic segmentation (a and b), obtained from the use of both strategies at the same time: data augmentation of the less representative classes and CLAHE method.

6.4.3 Discussion

The main goal of this work was to evaluate if there are approaches commonly used in outdoor environments to improve visual segmentation results for the underwater context and how much they can improve the overall performance. Therefore, some strategies were tested, such as approaches to improve image quality and data augmentation.

The main conclusions on the results are therefore the following:

- The better the quality of the visual images, the better the segmentation results are likely to be. Therefore, in the first experiments with 20 previously selected heterogeneous images, it was found that CLAHE and dive corrections improved the results for 3 out of 4 metrics. However, some inconsistencies were found when looking at individual cases, so it cannot be said with certainty which enhancement method is best suited to achieve better segmentation results.

- In the first test of the segmentation results with the different image enhancement approaches, we see that the CLAHE method is the only one that increases the mean IoU value, albeit with a small increase. Regarding the individual classes, this implementation improves the result of the coral, wreck and fish classes. The visual results confirm the quantitative results and for this reason this method was chosen as the best to improve the image quality, as it improves the visibility of the object boundaries and some important features that allow a better training of the model.

- In the first test of the data augmentation, the results with the dataset using the data augmentation increase the original result of the mean IoU by 2.3%, which shows that it is relevant and allows learning some classes with more diversification. There are 5 classes that improve their results, namely classes that are more difficult to learn, such as the Robot, because it is a class with fewer images in the original dataset (only 104 images), which increases its metric by 17%. However, this class and the Plants are the classes where increasing the number of images also leads to an unbalanced dataset.

- When a data augmentation was performed with 7 transformations for the images with plants, robots and wrecks and only 2 transformations for the rest, a slightly more balanced dataset was obtained and the mean IoU value increased to 54.4%, with the robots improving their result and the plants having a IoU value of 23.5% for the best model obtained during training.

- With the augmentation strategy for one class at a time, the IoU value obtained for the robot increases to a value of 41.4% (a value

never achieved before for this class). When analysing the plants, we see that the class initially has higher values than 0% in the test with the data augmentation. The values obtained are always higher than in the original test and reach a maximum of 25%. This lower value can be explained by the fact that the images with different variations are not very representative and for this reason it is a more difficult class to learn.

- The last test has the main goal to see if the technique of data augmentation and enhancement provides good results. The results obtained are very good, with a mean IoU value of 56% and an increased overall accuracy value of 81.8%. The mean accuracy value also increases, which means that the results are better in different classes. This is not the case for plants, which could mean that the enhancement confuses the model in the regions of this class (e.g. by setting many saliency features) and therefore this class remains the most difficult in the learning process, but in the visual results this class already appears more.

If we look at the last result with both strategies, we can conclude that all the changes made during the study are beneficial and that the results at this stage are better than in the original cases, because now there are 5 classes with more than 60% and only one with less than 30% for the IoU value, and therefore it is a more reliable result and can be used more easily in real contexts. It is important to mention that the results are different from the original results obtained with a higher training. This is a good result because it is to be expected that with longer training the results of some classes will be better as they are more difficult to train. Both strategies are therefore important in the context of semantic segmentation, especially in the underwater domain, where it is very difficult to obtain relevant images in the quantity required for deep learning techniques.

6.5 SUMMARY

Along this chapter, the relevance of the topic of object recognition for various tasks can be established, as it is an attractive field in the area of image understanding. It enables machines, such as autonomous vehicles, to interpret and understand information in a similar way to humans. Therefore, in most cases it is advantageous to rely on visual sensors as they capture a wealth of information and the information they provide is suitable for human interpretation in real time, making them more intuitive than other sensors. In underwater environments, they face some problems, such as sensitivity to lightning fluctuations,

limited visibility or scattering problems. Nevertheless, their use is interesting, especially for tasks involving the detection of objects.

As the exploration of the oceans is far from complete, this area needs to be explored further. Recently, deep learning has proven to be a powerful tool to overcome the shortcomings of traditional image classification methods by extracting unique, invariant and robust features. In the first study, a dataset previously developed for this context was used. The performance of the developed method proved to be good enough to correctly recognise the objects with an accuracy of 75% under most conditions. There are some classes where recognition was more difficult, such as pipelines or anchors, which can be confused with the background due to their colour. In the final experiment, it was found that the developed method can be used in a preliminary form in many situations, as the training was carried out with heterogeneous images. In this way, the model learns the features that can be used to better discriminate the objects, also in comparison with others in the literature. It was also important to mention that the developed method allows to perform the training with less than twice the time, which can be crucial to introduce some classes or features in the process.

As robots become more autonomous, it is increasingly important that they have a good representation of the scenario. This therefore requires an object recognition system that can not only identify an object, but also its shape and other characteristics such as size or shape. Visual object segmentation therefore presents itself as the precise delineation of individual objects at the pixel level, i.e. a dense classification. There are some approaches in the literature that are up to this task, but they have only been tested for outdoor images. Therefore, it was necessary to look for a dataset that allows this comparison, i.e. that contains the mask that serves as ground truth when evaluating the different methods. The SUIM dataset used proves to be the most complete and the results suggest that different methods from the literature are suitable for this semantic segmentation. Although the results are not perfect, it is worth mentioning that all methods provide good results for this real-world context, but when all measures and visual results are analysed and taken into account, SegNet stands out. This method has a higher capacity in the learning process and was therefore superior in terms of time and performance. This dataset has an unbalanced distribution of image examples per class. For this reason, there are two classes that are difficult for the model to learn and for which the results are poor. In the first experiment, an overall accuracy of about 80% was obtained, but with a mean IoU of only 53%, which shows that for some classes the ratio between the overlap area between the prediction and the real observation and the overlap area between the two (the best metric to evaluate this type of task) is lower.

Although the results are not bad in real contexts, it could be dangerous to work with them in a real application. For this reason, some strategies need to be tested to see if it is possible to improve these results, either by improving the quality of the images or by data augmentation. The better the quality of the visual images, the better the segmentation results are likely to be. However, we see that the CLAHE method is the only one that increases the mean IoU value, albeit with a small increase. Regarding the individual classes, the result of the coral, wreck and fish classes has improved, namely because the visibility of the object boundaries has been improved and some important features that allow better training of the model have been emphasised. The results with the dataset using the data augmentation increase the original result and show that learning some classes with more diversification is possible. There are 5 classes that improve their results, which is crucial. In the last experiment, combining the image quality improvement strategies with data augmentation, the results obtained are very good, with a mean IoU value of 56% and an increased overall accuracy of about 82%. It can be seen that the mean accuracy value also increases, which means that the results are better in different classes. All the changes made are favourable, because in the final result there are 5 classes with more than 60% and only one with less than 30% for the IoU value, and therefore it is a more reliable result and can be more easily used in real contexts. Both strategies are therefore important in the context of semantic segmentation, especially in the underwater domain, where it is very difficult to obtain relevant images in the quantity required for deep learning techniques. Therefore, it is believed that it is now easier to overcome this challenge in this environment and context, thus contributing to the advancement of autonomous vehicle perception in these scenarios.

Part III

FINAL CONSIDERATIONS

7 CONCLUSION AND FUTURE WORK

The thesis deals with the development of the reconstruction of scenes based on acoustic data and the identification, classification and segmentation of objects from visual images. The following conclusion can be drawn from this:

- **Experimental setup**

  For underwater environments, there is currently a lack of suitable data sets for real-world conditions compared to other areas. At the same time, there are some limitations in the available datasets, as they do not contain all the situations needed for the tests or the sensor configuration is different than expected, etc. Therefore, it is often necessary to use simulators to validate the developed methods. These can replicate some conditions of the intended scenarios, namely some difficulties associated with the underwater environment, using different sensors depending on their availability in the particular simulator. Therefore, some scenarios of the harbour bottom are replicated and the FLS data can be collected using the Stonefish simulator. After the simulation, the behaviour of the methods must be verified in real contexts and with data from real sensors. Therefore, it was necessary to select the best sensors for the different applications and combine them in a perception system. This task allows the collection of information about the environment with a camera, a FLS and a positioning system to reconstruct the mission performance of the robot.

- **Tri-dimensional reconstruction of scenes**

  A simple method has been developed to obtain a 3D reconstruction of the scenes under the sensor using an external source localisation (through an AruCo marker detection). This method offers two different filtering techniques: one according to the returned intensity and another according to the scan angle (which can be switched off). Moreover, unlike the usual techniques, only the central part of the 35° arc of the sensor is considered occupied (about 20°). As for the experiments with different materials: in order to choose a single value for all materials so that their visibility is not affected by the filtering techniques, a value of 50% of gain was considered. The method makes it possible to image smaller objects (without filtering techniques for the scan

angle) with a maximum error of 5 cm, which is a good result for real applications. For larger objects, the filter is activated and reduces the number of total points. The z dimensions of the objects are correct and the MSIS shows that it is suitable for this type of survey where the main goal is to find the obstacles near the sensor. On the other hand, the configuration of the scan angle can also influence the quality of the final map, because the larger the FOV, the slower the rotation speed of the sensor head. Finally, performing a discretisation process with different step sizes shows that it is possible to achieve a simpler representation of the scene, able to collect some information about the objects and with lower storage costs. This process is crucial, for example, in cases where the data is collected and sent to a centralised system without sacrificing storage space.

- **Acoustic Mosaicking**

  A mosaic method for FLS images was created and evaluated based on the same assumptions as for the visual approaches. In the experiments, two different binary extractors were compared and the results show that the ORB extractor provides a higher number of keypoints and consequently achieves the quality in the ratio test. One of the most important factors for the success of the mosaic is that the consecutive images have an overlap of more than 40%. On this basis, the result of combining different images is good, indicating that the images have been correctly merged based on their appearance. The motion estimated by homography is compared with ground truth and it is clear that it is well estimated. Finally, the importance of good matches is shown, because with lower values the mosaics are poor and the objects appearing in the image are often not correctly represented. A task that is very important in this environment is the automatic detection of objects that might appear during the exploration. In the case of objects such as a carabiner, a part of a tyre, a fine grid and a rope, the results of the mosaic are well combined and the final mosaics show quality in terms of visual and easy perception of what is seen in the image. In the case of the anchor, the plastic cup and the box, which have fewer key points in their body, it was necessary to make more parameterisations in order to increase the number of matches obtained and at the same time the quality of the final mosaic.

- **Visual Object Classification**

  A CNN was presented and its evaluation was performed under different conditions. A dataset with 7 man-made critical objects was developed. The results obtained show that the performance

is good enough to recognise the objects with an accuracy of 75% under most conditions. However, there are some classes where this task is more difficult. The object classification without these classes improves to an VA of 84%. The results obtained with the ResNet without pre-trained weights (ResNet_v1) can be compared with the developed method and show that it is better than the developed method, but it involves higher computational costs and requires more processing time. In general, the developed method can be used in many situations as the training was performed with heterogeneous images. In this way, the model learns the features that can be used to better distinguish the objects.

- **Visual Semantic Segmentation**

  A study to evaluate visual segmentation methods that have already proven their good performance in the literature (but were only tested under outdoor conditions). The SUIM dataset used proves to be complete to perform the comparison, even if it presents some problems common to this type of datasets. Although the results are not perfect, it is worth mentioning that all methods provide good results for this real-world context, but when all measurements and visual results are analysed and taken into account, SegNet stands out. Thus, the final results reach an overall accuracy of 80% and the mean value of IoU reaches 53%, even with very difficult test images. If we look at the results of the classes, we can see that there are two classes that are difficult for the model to learn and that the results are poor (less than 30% for robots and less than 10% for plants), which is dangerous for application in real contexts. This fact can be explained by the unbalanced dataset, which contains only a few images of these classes with a low variety of perspectives and size representations.

- **Improving the performance of Visual Semantic Segmentation**

  It was investigated whether there are approaches commonly used in outdoor environments that can be applied in visual segmentation methods for the underwater context and are able to improve the performance of semantic segmentation methods. Therefore, image quality enhancement and data augmentation approaches were applied to the dataset.

  In the first experiment with the different image enhancement approaches, the CLAHE method proved to be the only one that increased the mean IoU value, albeit with a small increase. For this reason, this method was selected as the best for improving image quality.

The results with the dataset using the data augmentation increase the original result of the mean IoU value by 2.3%, which shows that it is relevant and allows learning some classes with more diversification. However, the poor results for the Robot and Plant classes persist, as the training images also lead to an unbalanced dataset. When we perform a data augmentation for the images with plants, robots and wrecks, we obtain a slightly more balanced dataset and the performance obtained for the mean IoU value is 54.4%, with the robots improving their result and the plants showing a IoU value of 23.5% for the best model obtained during training. With the augmentation strategy for one class at a time, the IoU value obtained for the robot increases to a value of 41.4% (a value never before achieved for this class). When analysing the plants, the performance reaches a maximum of 25%. This lower value can be explained by the fact that the images with different variations are not very representative and for this reason it is a more difficult class to learn. The main objective of the last test is to find out if the technique of data augmentation and enhancement gives good results when applied simultaneously. The results obtained are very good, with a mean IoU value of 56% and an increased overall accuracy value of 81.8%. If we consider the result with both strategies, we can conclude that all the changes made during the study are beneficial and that the results at this stage are better than in the original cases, because now there are 5 classes with more than 60% and only one with less than 30% for the IoU value, and therefore it is a more reliable result and can be used more easily in real contexts. Both strategies are therefore important in the context of semantic segmentation, especially in the underwater domain, where it is very difficult to obtain relevant images in the quantity required for deep learning techniques.

Future Work

The results presented in this thesis show good progress in both areas: scene reconstruction and object recognition, which improves the robot's perception of the mission scenario and allows it to better explore the underwater environment. Nevertheless, some steps can be taken to improve the approaches:

- To obtain new real data to better test the different parts of this work under different visibility conditions and ensure the feasibility and accuracy of the results.
- Analyse the behaviour of the classification method using acoustic data to be able to apply it in all visibility conditions, as the

cameras do not work well in cases without lighting or with high turbidity.

- Extend the classification and segmentation problem to areas that are of great importance nowadays, such as the classification of waste, since it is a major problem for our oceans and waters and many campaigns have been carried out to recover it in order to reduce the level of pollution.
- Progress in semantic mapping, which is an important task for the different fields of robotics. Nowadays, it is crucial for robots in underwater research to map the environment and recognise the elements they found in this map. In this way, they can react accordingly. In this particular case, it is interesting to use the representation of the scenario obtained by the mosaic technique to classify the objects encountered resorting to optical or acoustic images.

BIBLIOGRAPHY

[1] H. Lu, Y. Li, and S. Serikawa. "Computer Vision for Ocean Observing". In: *Artificial Intelligence and Computer Vision*. Ed. by H. Lu and Y. Li. Cham: Springer International Publishing, 2017, pp. 1–16. ISBN: 978-3-319-46245-5. DOI: 10.1007/978-3-319-46245-5_1 (cit. on p. 9).

[2] J. Henderson, O. Pizarro, M. Johnson-Roberson, and I. Mahon. "Mapping Submerged Archaeological Sites using Stereo-Vision Photogrammetry". In: *International Journal of Nautical Archaeology* 42.2 (2013), pp. 243–256. DOI: https://doi.org/10.1111/1095-9270.12016. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1095-9270.12016 (cit. on p. 10).

[3] M. Roser, M. Dunbabin, and A. Geiger. "Simultaneous underwater visibility assessment, enhancement and improved stereo". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 2014, pp. 3840–3847. DOI: 10.1109/ICRA.2014.6907416 (cit. on p. 10).

[4] H. Lu, Y. Li, Y. Zhang, M. Chen, S. Serikawa, and H. Kim. "Underwater Optical Image Processing: A Comprehensive Review". In: *CoRR* abs/1702.03600 (2017). arXiv: 1702.03600 (cit. on p. 10).

[5] E. Galceran, V. Djapic, M. Carreras, and D. P. Williams. "A Real-time Underwater Object Detection Algorithm for Multi-beam Forward Looking Sonar". In: *IFAC Proceedings Volumes* 45.5 (2012). 3rd IFAC Workshop on Navigation, Guidance and Control of Underwater Vehicles, pp. 306–311. ISSN: 1474-6670. DOI: https://doi.org/10.3182/20120410-3-PT-4028.00051 (cit. on pp. 10, 25).

[6] Z. Chen, Z. Zhang, F. Dai, Y. Bu, and H. Wang. "Monocular Vision-Based Underwater Object Detection". In: *Sensors* 17.8 (2017). ISSN: 1424-8220. DOI: 10.3390/s17081784 (cit. on pp. 10, 25).

[7] M. Kido, Y. Osada, and H. Fujimoto. "Temporal variation of sound speed in ocean: A comparison between GPS/acoustic and in situ measurements". In: *Earth, Planets and Space* 60 (Mar. 2008). DOI: 10.1186/BF03352785 (cit. on p. 10).

[8] A. Umanath and S. Malarkkan. "Underwater object identification and recognition with sonar images using soft computing techniques". In: *Indian Journal of Geo-Marine Sciences* 47 (Mar. 2018), pp. 665–673 (cit. on pp. 11, 29).

[9] M. Massot-Campos and G. Oliver-Codina. "Optical Sensors and Methods for Underwater 3D Reconstruction". In: *Sensors* 15.12 (2015), pp. 31525–31557. ISSN: 1424-8220. DOI: 10.3390/s151229864 (cit. on p. 11).

[10] C. Sukanya, R. Gokul, and V. Paul. "A Survey on Object Recognition Methods". In: *International Journal of Computer Science Engineering and Technology* 6.1 (2016), pp. 48–52 (cit. on pp. 12, 24).

[11] Y. LeCun, K. Kavukcuoglu, and C. Farabet. "Convolutional networks and applications in vision". In: *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. 2010, pp. 253–256. DOI: 10.1109/ISCAS.2010.5537907 (cit. on pp. 13, 27).

[12] Y. LeCun, Y. Bengio, and G. Hinton. "Deep Learning". In: *Nature* 521 (May 2015), pp. 436–44. DOI: 10.1038/nature14539 (cit. on p. 14).

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Commun. ACM* 60.6 (2017), 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386 (cit. on p. 14).

[14] Y. Guo, G. Nie, W. Gao, and M. Liao. "2D Semantic Segmentation: Recent Developments and Future Directions". In: *Future Internet* 15.6 (2023). ISSN: 1999-5903. DOI: 10.3390/fi15060205 (cit. on p. 15).

[15] N. Sünderhauf, F. Dayoub, S. McMahon, B. Talbot, R. Schulz, P. Corke, G. F. Wyeth, B. Upcroft, and M. Milford. "Place categorization and semantic mapping on a mobile robot". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)* (2015), pp. 5729–5736 (cit. on pp. 15, 32).

[16] J. Cheng, L. Zhang, Q. Chen, X. Hu, and J. Cai. "A review of visual SLAM methods for autonomous driving vehicles". In: *Engineering Applications of Artificial Intelligence* 114 (2022), p. 104992. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2022.104992 (cit. on p. 16).

[17] K. Pathak, A. Birk, and N. Vaskevicius. "Plane-based registration of sonar data for underwater 3D mapping". In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2010, pp. 4880–4885. DOI: 10.1109/IROS.2010.5650953 (cit. on p. 18).

[18] H. Bullow and A. Birk. "Spectral registration of noisy sonar data for underwater 3D Mapping". In: *Auton. Robots* 30 (Apr. 2011), pp. 307–331. DOI: 10.1007/s10514-011-9221-8 (cit. on p. 18).

[19] Z. Shen, J. Song, K. Mittal, and S. Gupta. "An autonomous integrated system for 3-D underwater terrain map reconstruction". In: *OCEANS 2016 MTS/IEEE Monterey*. 2016, pp. 1–6. DOI: 10.1109/OCEANS.2016.7761067 (cit. on p. 18).

[20] S. Young-eun and C. Seung-Joon. "Underwater 3D Reconstruction for Underwater Construction Robot Based on 2D Multibeam Imaging Sonar". In: *J. Ocean Eng. Technol.* 30.3 (2016), pp. 227–233. DOI: 10.5574/KSOE.2016.30.3.227. eprint: http://www.joet.org/journal/view.php?number=103 (cit. on p. 18).

[21] T. Guerneve and Y. R. Pétillot. "3D Reconstruction of Underwater objects using wide-aperture Imaging SONAR". In: vol. 35. 6. 2018, pp. 890–905. DOI: 10.1002/rob.21783 (cit. on p. 19).

[22] A. Mallios, P. Ridao, D. Ribas, M. Carreras, and R. Camilli. "Toward Autonomous Exploration in Confined Underwater Environments". In: *Journal of Field Robotics* 33.7 (2016), pp. 994–1012. DOI: https://doi.org/10.1002/rob.21640. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21640 (cit. on p. 19).

[23] S. Kwon, J. Park, and J. Kim. "3D reconstruction of underwater objects using a wide-beam imaging sonar". In: *2017 IEEE Underwater Technology (UT)*. 2017, pp. 1–4. DOI: 10.1109/UT.2017.7890306 (cit. on p. 20).

[24] H. Joe, J. Kim, and S.-C. Yu. "Sensor Fusion-based 3D Reconstruction by Two Sonar Devices for Seabed Mapping". In: *IFAC-PapersOnLine* 52.21 (2019). 12th IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS 2019, pp. 169–174. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2019.12.302 (cit. on p. 20).

[25] J. Wang, S. Bai, and B. Englot. "Underwater localization and 3D mapping of submerged structures with a single-beam scanning sonar". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017, pp. 4898–4905. DOI: 10.1109/ICRA.2017.7989567 (cit. on p. 20).

[26] J. Henderson, O. Pizarro, M. Johnson-Roberson, and I. Mahon. "Mapping Submerged Archaeological Sites using Stereo-Vision Photogrammetry". In: *International Journal of Nautical Archaeology* 42.2 (2013), pp. 243–256. DOI: https://doi.org/10.1111/1095-9270.12016. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1095-9270.12016 (cit. on p. 20).

[27] N. Weidner, S. Rahman, A. Q. Li, and I. Rekleitis. "Underwater cave mapping using stereo vision". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017, pp. 5709–5715. DOI: 10.1109/ICRA.2017.7989672 (cit. on p. 21).

[28] S. Negahdaripour, H. Sekkati, and H. Pirsiavash. "Opti-Acoustic Stereo Imaging, System Calibration and 3-D Reconstruction". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383361 (cit. on p. 21).

[29] M. Hurtós, X. C. i Soler, and J. Salvi. "Integration of optical and acoustic sensors for 3D underwater scene reconstruction." In: *Instrumentation viewpoint* (2009), pp. 43–43 (cit. on p. 22).

[30] A. P. Nunes, A. R. S. Gaspar, A. M. Pinto, and A. C. Matos. "A mosaicking technique for object identification in underwater environments". In: *Sensor Review* 39.3 (2019), pp. 387–396. DOI: https://doi.org/10.1108/SR-04-2018-0089 (cit. on pp. 22, 81).

[31] D. Vaghela and P. K. Naina. *A Review of Image Mosaicing Techniques*. 2014. arXiv: 1405.2539 [cs.CV] (cit. on p. 22).

[32] A. Pandey and U. C. Pati. "Image mosaicing: A deeper insight". In: *Image and Vision Computing* 89 (2019), pp. 236–257. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2019.07.002 (cit. on p. 22).

[33] S. Negahdaripour, P. Firoozfam, and P. Sabzmeydani. "On processing and registration of forward-scan acoustic video imagery". In: *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*. 2005, pp. 452–459. DOI: 10.1109/CRV.2005.57 (cit. on p. 23).

[34] S. Song, J. M. Herrmann, K. Liu, S. Li, and X. Feng. "Forward-looking sonar image mosaicking by feature tracking". In: *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. Qingdao, China: IEEE Press, 2016, 1613–1618. DOI: 10.1109/ROBIO.2016.7866558 (cit. on p. 23).

[35] N. Hurtós, D. Ribas, X. Cufí, Y. Petillot, and J. Salvi. "Fourier-based Registration for Robust Forward-looking Sonar Mosaicing in Low-visibility Underwater Environments". In: *Journal of Field Robotics* 32.1 (2015), pp. 123–151. DOI: https://doi.org/10.1002/rob.21516. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21516 (cit. on p. 23).

[36] K. Kim, N. Intrator, and N. Neretti. "Image registration and mosaicing of noisy acoustic camera images". In: Jan. 2005, pp. 527–530. ISBN: 0-7803-8715-5. DOI: 10.1109/ICECS.2004.1399734 (cit. on p. 23).

[37] S. Negahdaripour, M. D. Aykin, and S. Sinnarajah. "Dynamic scene analysis and mosaicing of benthic habitats by FS sonar imaging - Issues and complexities". In: *OCEANS'11 MTS/IEEE KONA*. 2011, pp. 1–7. DOI: 10.23919/OCEANS.2011.6107093 (cit. on p. 23).

[38] F. Ferreira, V. Djapic, M. Micheli, and M. Caccia. "Forward looking sonar mosaicing for Mine Countermeasures". In: *Annual Reviews in Control* 40 (2015), pp. 212–226. ISSN: 1367-5788. DOI: https://doi.org/10.1016/j.arcontrol.2015.09.014 (cit. on p. 23).

[39] F. Ferreira, V. Djapic, and M. Caccia. "Real-time Mosaicing of Large Scale Areas with Forward Looking Sonar". In: *IFAC-PapersOnLine* 48.2 (2015). 4th IFAC Workshop onNavigation, Guidance and Controlof Underwater VehiclesNGCUV 2015, pp. 32–37. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2015.06.006 (cit. on p. 24).

[40] D. L. Rizzini, F. Kallasi, F. Oleari, and S. Caselli. "Investigation of Vision-Based Underwater Object Detection with Multiple Datasets". In: *International Journal of Advanced Robotic Systems* 12.6 (2015), p. 77. DOI: 10.5772/60526 (cit. on p. 25).

[41] A. Pavin. "Underwater object recognition in photo images". In: *OCEANS 2015 - MTS/IEEE Washington*. 2015, pp. 1–6. DOI: 10.23919/OCEANS.2015.7401926 (cit. on p. 25).

[42] J. Tian, S. Xue, and H. Huang. "Classification of Underwater Objects Based on Probabilistic Neural Network". In: vol. 2. Jan. 2009, pp. 38–42. DOI: 10.1109/ICNC.2009.362 (cit. on p. 26).

[43] N. Thakur and D. B. Maheshwari. "A Review on Image Classification Approaches and Techniques". In: *International Research Journal of Engineering and Technology (IRJET)* 4.11 (2017) (cit. on p. 26).

[44] M. Moniruzzaman, S. Islam, M. Bennamoun, and P. Lavery. "Deep Learning on Underwater Marine Object Detection: A Survey". In: Nov. 2017, pp. 150–160. ISBN: 978-3-319-70352-7. DOI: 10.1007/978-3-319-70353-4_13 (cit. on p. 27).

[45] A. Salman, A. Jalal, F. Shafait, A. Mian, M. Shortis, J. Seager, and E. Harvey. "Fish species classification in unconstrained underwater environments based on deep learning". In: *Limnology and Oceanography: Methods* 14.9 (2016), pp. 570–585. DOI: https://doi.org/10.1002/lom3.10113 (cit. on p. 27).

[46] L. Jin and H. Liang. "Deep learning for underwater image recognition in small sample size situations". In: *OCEANS 2017 - Aberdeen*. 2017, pp. 1–4. DOI: 10.1109/OCEANSE.2017.8084645 (cit. on p. 27).

[47] S. A. Siddiqui, A. Salman, M. I. Malik, F. Shafait, A. Mian, M. R. Shortis, and E. S. Harvey. "Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data". In: *ICES Journal of Marine Science* 75.1 (July 2017), pp. 374–389. ISSN: 1054-3139. DOI: 10.1093/icesjms/fsx109 (cit. on p. 28).

[48] X. Wang, J. Ouyang, D. Li, and G. Zhang. "Underwater Object Recognition Based on Deep Encoding-Decoding Network". In: *Journal of Ocean University of China* 18 (2019), pp. 376–382 (cit. on p. 28).

[49] L. R. Fuchs, A. Gällström, and J. Folkesson. "Object Recognition in Forward Looking Sonar Images using Transfer Learning". In: *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*. 2018, pp. 1–6. DOI: 10.1109/AUV.2018.8729686 (cit. on p. 28).

[50] W. Chen, C. He, C. Ji, M. Zhang, and S. Chen. "An improved K-means algorithm for underwater image background segmentation". In: *Multimedia Tools and Applications* 80 (June 2021). DOI: 10.1007/s11042-021-10693-7 (cit. on p. 30).

[51] A. Garcia-Garcia, S. Orts, S. Oprea, V. Villena-Martinez, and J. G. Rodríguez. "A Review on Deep Learning Techniques Applied to Semantic Segmentation". In: *ArXiv* abs/1704.06857 (2017) (cit. on p. 30).

[52] Y. Zhou, J. Wang, B. Li, Q. Meng, E. Rocco, and A. Saiani. "Underwater scene segmentation by deep neural network". In: (Mar. 2019) (cit. on p. 30).

[53] V. Badrinarayanan, A. Kendall, and R. Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), pp. 2481–2495. DOI: 10.1109/TPAMI.2016.2644615 (cit. on pp. 31, 110).

[54] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. "Rethinking Atrous Convolution for Semantic Image Segmentation". In: *ArXiv* abs/1706.05587 (2017) (cit. on p. 31).

[55] P. Drews-Jr, I. Souza, I. Maurell, E. Protas, and S. Botelho. "Underwater image segmentation in the wild using deep learning". In: *Journal of the Brazilian Computer Society* 27 (Dec. 2021). DOI: 10.1186/s13173-021-00117-7 (cit. on p. 31).

[56] J. Wang, X. He, F. Shao, G. Lu, R. Hu, and Q. Jiang. "Semantic segmentation method of underwater images based on encoder-decoder architecture". In: *PLOS ONE* 17.8 (Aug. 2022), pp. 1–19. DOI: 10.1371/journal.pone.0272666 (cit. on p. 31).

[57] K. H. Sanila, A. A. Balakrishnan, and M. H. Supriya. "Underwater Image Enhancement Using White Balance, USM and CLHE". In: *2019 International Symposium on Ocean Technology (SYMPOL)*. 2019, pp. 106–116. DOI: 10.1109/SYMPOL48207.2019.9005301 (cit. on p. 31).

[58] K. Alomar, H. I. Aysel, and X. Cai. "Data Augmentation in Classification and Segmentation: A Survey and New Strategies". In: *Journal of Imaging* 9 (2 Feb. 2023). ISSN: 2313433X. DOI: 10.3390/jimaging9020046 (cit. on p. 31).

[59] M. Jian, X. Liu, H. Luo, X. Lu, H. Yu, and J. Dong. "Underwater image processing and analysis: A review". In: *Signal Processing: Image Communication* 91 (Feb. 2021). ISSN: 09235965. DOI: 10.1016/j.image.2020.116088 (cit. on p. 31).

[60] N. Hurtós, N. Palomeras, A. Carrera, and M. Carreras. "Autonomous detection, following and mapping of an underwater chain using sonar". In: *Ocean Engineering* 130 (2017), pp. 336–350. ISSN: 0029-8018. DOI: https://doi.org/10.1016/j.oceaneng.2016.11.072 (cit. on p. 32).

[61] N. Sunderhauf, T. Pham, Y. Latif, M. Milford, and I. Reid. "Meaningful Maps with object oriented Semantic Mapping". In: Sept. 2017, pp. 5079–5085. DOI: 10.1109/IROS.2017.8206392 (cit. on p. 33).

[62] Z. Qian, K. Patath, J. Fu, and J. Xiao. "Semantic SLAM with Autonomous Object-Level Data Association". In: May 2021, pp. 11203–11209. DOI: 10.1109/ICRA48506.2021.9561532 (cit. on p. 33).

[63] H. Bavle, P. De La Puente, J. P. How, and P. Campoy. "VPS-SLAM: Visual Planar Semantic SLAM for Aerial Robotic Systems". In: *IEEE Access* 8 (2020), pp. 60704–60718. DOI: 10.1109/ACCESS.2020.2983121 (cit. on p. 33).

[64] S. L. Bowman, K. Daniilidis, and G. J. Pappas. "Robust and Efficient Semantic SLAM with Semantic Keypoints". In: 2020 (cit. on p. 33).

[65] M. Machado, P. Drews, P. Núñez, and S. Botelho. "Semantic Mapping on Underwater Environment Using Sonar Data". In: *2016 XIII Latin American Robotics Symposium and IV Brazilian Robotics Symposium (LARS/SBR)*. 2016, pp. 245–250. DOI: 10.1109/LARS-SBR.2016.48 (cit. on p. 34).

[66] M. Dos Santos, P. O. Ribeiro, P. Núñez, P. Drews-Jr, and S. Botelho. "Object Classification in Semi Structured Enviroment Using Forward-Looking Sonar". In: *Sensors* 17.10 (2017). ISSN: 1424-8220. DOI: 10.3390/s17102235 (cit. on p. 34).

[67] T. Guerneve, K. Subr, and Y. Petillot. "Underwater 3D structures as semantic landmarks in SONAR mapping". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017, pp. 614–619. DOI: 10.1109/IROS.2017.8202215 (cit. on p. 34).

[68] K. Himri, R. Pi, P. Ridao, N. Gracias, A. Palomer, and N. Palomeras. "Object Recognition and Pose Estimation using Laser scans For Advanced Underwater Manipulation". In: *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*. 2018, pp. 1–6. DOI: 10.1109/AUV.2018.8729742 (cit. on p. 34).

[69] K. Himri, P. Ridao, N. Gracias, A. Palomer, N. Palomeras, and R. Pi. "Semantic SLAM for an AUV using object recognition from point clouds". In: *IFAC-PapersOnLine* 51.29 (2018). 11th IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS 2018, pp. 360–365. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2018.09.497 (cit. on p. 34).

[70] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on p. 35).

[71] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman. "Automated annotation of coral reef survey images". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 1170–1177. DOI: 10.1109/CVPR.2012.6247798 (cit. on p. 35).

[72] O. Beijbom, T. Treibitz, D. Kline, G. Eyal, A. Khen, B. Neal, Y. Loya, B. Mitchell, and D. Kriegman. "Improving Automated Annotation of Benthic Survey Images Using Wide-band Fluorescence". In: *Scientific Reports* 6 (Mar. 2016). DOI: 10.1038/srep23166 (cit. on p. 35).

[73] I. Alonso, M. Yuval, G. Eyal, T. Treibitz, and A. C. Murillo. "CoralSeg: Learning Coral Segmentation from Sparse Annotations". In: *Journal of Field Robotics*. 2019 (cit. on p. 35).

[74] A. Saleh, I. H. Laradji, D. A. Konovalov, M. Bradley, D. Vázquez, and M. Sheaves. "A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis". In: *Scientific Reports* 10 (2020) (cit. on p. 35).

[75] N. Garcia-D'Urso, A. Galán Cuenca, P. Pérez-Sánchez, P. Climent i Pérez, A. Guilló, J. Azorin-Lopez, M. Saval-Calvo, J. Guillén-Nieto, and G. Soler-Capdepón. "The DeepFish computer vision dataset for fish instance segmentation, classification, and size estimation". In: *Scientific Data* 9 (June 2022), p. 287. DOI: 10.1038/s41597-022-01416-0 (cit. on p. 36).

[76] M. Jian, Q. Qi, H. Yu, J. Dong, C. Cui, X. Nie, H. Zhang, Y. Yin, and K.-M. Lam. "The extended marine underwater environment database and baseline evaluations". In: *Applied Soft Computing* 80 (2019), pp. 425–437. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2019.04.025 (cit. on p. 36).

[77] J. Hong, M. Fulton, and J. Sattar. "TrashCan: A Semantically-Segmented Dataset towards Visual Detection of Marine Debris". In: *CoRR* abs/2007.08097 (2020). arXiv: 2007.08097 (cit. on p. 36).

[78] M. J. Islam, P. Luo, and J. Sattar. "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception". In: *CoRR* abs/2002.01155 (2020). arXiv: 2002.01155 (cit. on p. 36).

[79] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar. "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020), pp. 1769–1776. DOI: https://doi.org/10.48550/arXiv.2004.01241 (cit. on pp. 37, 113).

[80] P. Cieślak. "Stonefish: An Advanced Open-Source Simulation Tool Designed for Marine Robotics, With a ROS Interface". In: *OCEANS 2019 - Marseille*. 2019. DOI: 10.1109/OCEANSE.2019.8867434 (cit. on p. 42).

[81] D. Ribas, N. Palomeras, P. Ridao, M. Carreras, and A. Mallios. "Girona 500 AUV: From Survey to Intervention". In: *IEEE/ASME Transactions on Mechatronics* 17.1 (2012), pp. 46–53. DOI: 10.1109/TMECH.2011.2174065 (cit. on p. 44).

[82] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson. "Pose-Graph SLAM Using Forward-Looking Sonar". In: *IEEE Robotics and Automation Letters* 3.3 (2018), pp. 2330–2337. DOI: 10.1109/LRA.2018.2809510 (cit. on p. 50).

[83] H. Zhang, M. Tian, G. Shao, J. Cheng, and J. Liu. "Target Detection of Forward-Looking Sonar Image Based on Improved YOLOv5". In: *IEEE Access* 10 (2022), pp. 18023–18034. DOI: 10.1109/ACCESS.2022.3150339 (cit. on p. 50).

[84] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard. "OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees". In: *Autonomous Robots* (2013). Software available at http://octomap.github.com. DOI: 10.1007/s10514-012-9321-0 (cit. on p. 68).

[85] E. Hernández, P. Ridao, A. Mallios, and M. Carreras. "Occupancy grid mapping in an underwater structured environment". In: *IFAC Conference on Manoeuvring and Control of Marine Craft*. Vol. 42. 18. 2009, pp. 286–291. ISBN: 9783902661517. DOI: 10.3182/20090916-3-BR-3001.0049 (cit. on p. 68).

[86] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV] (cit. on p. 102).

[87] M. P. Shah. *Semantic Segmentation Architectures Implemented in PyTorch*. https://github.com/meetshah1995/pytorch-semseg. 2017 (cit. on p. 109).

[88] L. Jonathan, S. Evan, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. 2015. DOI: https://doi.org/10.48550/arXiv.1411.4038 (cit. on p. 110).

[89] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. "Pyramid Scene Parsing Network". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6230–6239. DOI: 10.1109/CVPR.2017.660 (cit. on p. 110).

[90] K. H. Sanila, A. A. Balakrishnan, and M. H. Supriya. "Underwater Image Enhancement Using White Balance, USM and CLHE". In: *2019 International Symposium on Ocean Technology (SYMPOL)*. 2019, pp. 106–116. DOI: 10.1109/SYMPOL48207.2019.9005301 (cit. on p. 127).

[91] S. Pizer, R. Johnston, J. Ericksen, B. Yankaskas, and K. Muller. "Contrast-limited adaptive histogram equalization: speed and effectiveness". In: *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing*. 1990, pp. 337–345. DOI: 10.1109/VBC.1990.109340 (cit. on p. 127).

[92] R. Ramanath and M. S. Drew. "White Balance". In: *Computer Vision: A Reference Guide*. Boston, MA: Springer US, 2014, pp. 885–888. ISBN: 978-0-387-31439-6. DOI: 10.1007/978-0-387-31439-6_467 (cit. on p. 127).

[93] M. Afifi and M. S. Brown. "What Else Can Fool Deep Learning? Addressing Color Constancy Errors on Deep Neural Network Performance". In: *The IEEE International Conference on Computer Vision (ICCV)*. 2019. DOI: 10.1109/ICCV.2019.00033 (cit. on p. 127).

[94] A. R. Reibman, R. M. Bell, and S. Gray. "Quality assessment for super-resolution image enhancement". In: *2006 International Conference on Image Processing*. 2006, pp. 2017–2020. DOI: 10.1109/ICIP.2006.312895 (cit. on p. 129).

[95] W. Zhou, Z. Wang, and Z. Chen. "Image super-resolution quality assessment: Structural fidelity versus statistical naturalness". In: *13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2021, pp. 61–64. DOI: 10.48550/arXiv.2105.07139 (cit. on p. 129).

[96] A. Mittal, A. K. Moorthy, and A. C. Bovik. "No-Reference Image Quality Assessment in the Spatial Domain". In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708. DOI: 10.1109/TIP.2012.2214050 (cit. on p. 129).

[97] V. N, P. D, M. C. Bh, S. S. Channappayya, and S. S. Medasani. "Blind image quality evaluation using perception based features". In: *2015 Twenty First National Conference on Communications (NCC)*. 2015, pp. 1–6. DOI: 10.1109/NCC.2015.7084843 (cit. on p. 130).

[98] A. Mittal, R. Soundararajan, and A. C. Bovik. "Making a "Completely Blind" Image Quality Analyzer". In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 209–212. DOI: 10.1109/LSP.2012.2227726 (cit. on p. 130).

[99] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang. "Learning a no-reference quality metric for single-image super-resolution". In: *Computer Vision and Image Understanding* 158 (2017), pp. 1–16. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2016.12.009 (cit. on p. 130).

[100] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. "Albumentations: Fast and Flexible Image Augmentations". In: *Information* 11.2 (2020), pp. 2078–2489. DOI: 10.3390/info11020125 (cit. on p. 132).