Journal of
Marine Science and Engineering

MDPI

Article

Improving Semantic Segmentation Performance in Underwater Images

Alexandra Nunes * and Aníbal Matos

Faculty of Engineering-University of Porto (FEUP), Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), 4200 Porto, Portugal
* Correspondence: apn@inesctec.pt

Abstract: Nowadays, semantic segmentation is used increasingly often in exploration by underwater robots. For example, it is used in autonomous navigation so that the robot can recognise the elements of its environment during the mission to avoid collisions. Other applications include the search for archaeological artefacts, the inspection of underwater structures or in species monitoring. Therefore, it is necessary to improve the performance in these tasks as much as possible. To this end, we compare some methods for image quality improvement and data augmentation and test whether higher performance metrics can be achieved with both strategies. The experiments are performed with the SegNet implementation and the SUIM dataset with eight common underwater classes to compare the obtained results with the already known ones. The results obtained with both strategies show that they are beneficial and lead to better performance results by achieving a mean IoU of 56% and an increased overall accuracy of 81.8%. The result for the individual classes shows that there are five classes with an IoU value close to 60% and only one class with an IoU value less than 30%, which is a more reliable result and is easier to use in real contexts.

Keywords: semantic segmentation; data augmentation; enhancement techniques; underwater; visual information

Citation: Nunes, A.; Matos, A. Improving Semantic Segmentation Performance in Underwater Images. *J. Mar. Sci. Eng.* **2023**, *11*, 2268. https://doi.org/10.3390/jmse11122268

Academic Editors: Anna Nora Tassetti, Adriano Mancini and Pierluigi Penna

Received: 15 October 2023
Revised: 23 November 2023
Accepted: 27 November 2023
Published: 29 November 2023

1. Introduction

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 during underwater exploration, especially for autonomous navigation when the robot needs to recognise the elements it encounters during the mission; for example, to avoid collisions. Likewise, it can 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. Further 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. The main objective of this task is to assign each pixel of an image to a corresponding class of the represented image. This is a dense classification and it is often difficult to obtain good prediction results. In traditional approaches, there are some problems with the accuracy of the results because the data obtained in underwater environments have challenges such as colour distortion, low contrast, noise or uneven illumination, and for these reasons, some important information is lost. In addition, traditional methods are generally not very transferable or robust, so the segmentation result of a single traditional method is poor in most cases [1]. It is therefore necessary to resort to advanced approaches, often involving Deep Learning [2], to better address these underwater challenges [3–5]. For these methods to work well for this type of task, which requires training models, appropriate 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 to consider for different contexts or classes [6–9]. When there are suitable

datasets for the general underwater context, they often do not have masks with the ground truth [10], which is a challenge when they need to be manually labelled. In a previous study by the authors, the SUIM dataset [11] was chosen to comparatively test different machine learning approaches (SegNet [12], Pyramid Scene Parsing Network—PSPNET [13] and two versions of Fully Convolutional Networks—FCNN [14], which are part of a Pytorch repository [15]) for semantic segmentation. This dataset was chosen because it contains seven common classes of the underwater world with different perspectives and visual conditions. In this research [16], the Segnet implementation, a deep, fully convolutional neural network architecture for pixel-level semantic segmentation that is efficient in terms of both memory and computation time, stands out in terms of key performance metrics. Thus, the overall accuracy, mean accuracy and mean Intersection over Union (IoU) reach 80%, 64% and 52%, respectively, but do not perform as well when we analyse the results of the individual classes. Only three of the overall classes achieve more than 60% and two classes less than 30%, with a large training set for 100 evaluations. Although the results obtained are not bad in real contexts, it could be dangerous to work with them during a real mission. Thus, 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. [17,18]. The first goal was to find the right parameters, and this was achieved in the first iteration. Now, additional strategies need to be tested to see if they are successful in the underwater environment.

Thus, the main purpose of this work was to find out if there are some approaches to augmenting the quality and diversity of training data commonly used in outdoor environments to improve the results of segmentation in the underwater context (see Figure 1).

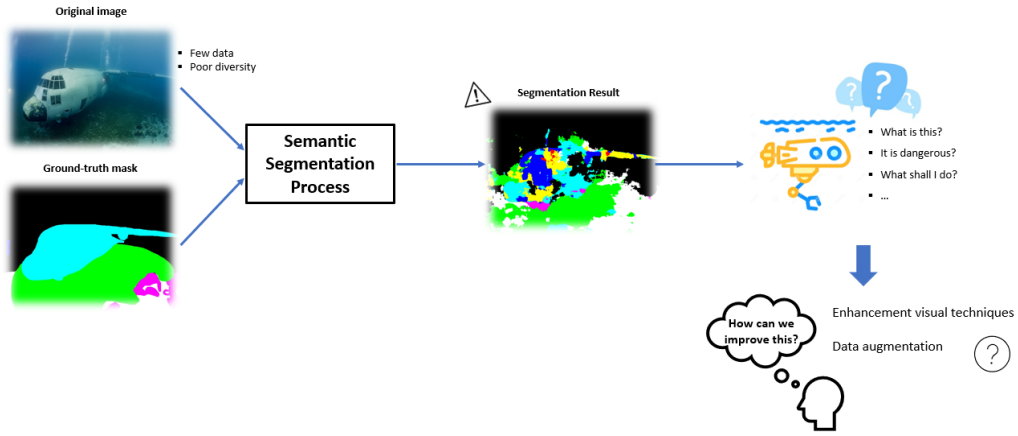

Figure 1. Diagram of the general framework of the proposal.

The most important contributions of this work are:

- The evaluation of the impact of the enhancement methods on semantic segmentation results;
- The exploration of data augmentation strategies that allow to increase the number of training images and thus the diversity of the original datasets;
- The robustness of the semantic segmentation result by using both strategies: Data augmentation and visual technique enhancement.

This paper is arranged as follows: Section 2 describes the background of this work with the explanation of the previously obtained results, i.e., the starting point of this study. Section 3 presents the main methods used to improve the quality of the underwater images and some approaches to expand the number of samples in a dataset. Next, the Section 4 shows the main findings in terms of accuracy and visual results and some discussion about them, and finally, Section 5 presents the main conclusions and some ideas for future work.

2. Background

In semantic segmentation, all objects of the same class are given the same label. When used with machine learning, the system is trained with already segmented datasets to accurately identify elements and segment an unknown set of images. As mentioned earlier, in an earlier work, the authors conducted a detailed comparative study of four different implementations for semantic segmentation in an underwater context [16]. To achieve this, after reviewing the available online data, we used the SUIM dataset, which is one of the most complete and could be useful, although it does not yet contain all the intended objects. It is not a large dataset, as it is not easy to obtain underwater images for these scenarios, but it provides different perspectives, as well as colour and size information. Another important aspect of this dataset is that it contains the ground truth for the different images, i.e., masks with the pixels labelled according to the classes. There are eight 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 Seabed and Rocks (Class 7). However, each class has a different number of samples, which leads to an unbalanced dataset and could be a problem in the training phase. Figure 2 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.

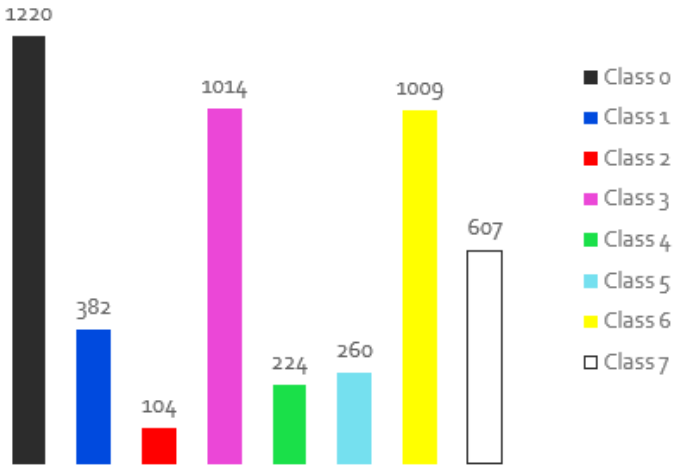

Figure 2. Distribution of the individual classes for the images of the original training set.

Therefore, a series of experiments was conducted under different conditions to find out whether SegNet, PSPNET or FCNNs is best suited for the intended context. Segnet stands out in the results obtained. This approach is a deep convolutional encoder–decoder architecture for image segmentation, which has good performance in terms of memory and computation time. It also offers a smaller number of trainable parameters than other approaches. The study carried out using the selected underwater dataset showed a good trade-off between the accuracy of the results and the time required, even though time is not the most important factor. In order to test this type of application, it is important to resort to different techniques to measure the performance obtained after the training. Therefore, in addition to observing the final results, it is usually possible to use the overall and mean accuracy, which refers to the number of correctly assigned pixels (the mean takes into account the predicted classes). It is also possible to calculate the IoU, which calculates the performance of each class, taking into account the area of overlap between the prediction and the real observation and the union of the two areas.

An example of the final results obtained is presented using a test with the original dataset with more than 1500 images for training and 110 for evaluation. It is important to note that the test dataset is hard, as it contains the main challenges encountered in underwater visual data collection such as lighting inconsistencies, boundaries of objects

not perceived as a whole, etc. The test is conducted after every 500 random images, and 100 evaluations were conducted to obtain the final model and performance metrics. However, the best model was found after 45,000 images, i.e., after 90 evaluations. Table 1 summarises the results in terms of overall accuracy, mean accuracy, mean IoU and the values for this metric for each class.

Table 1. Performance measurements [%] for the best result obtained with the original dataset in a total of 100 evaluations.

Overall Acc	Mean Acc	Mean IoU	IoU 0	IoU 1	IoU 2	IoU 3	IoU 4	IoU 5	IoU 6	IoU 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 occur because these classes have a smaller number of images for training compared to the other classes. In Figure 3, the results of semantic segmentation for five examples can be seen. In general, it was found that the boundaries of the elements were correctly selected, but some of them were assigned to the wrong class; see the diver in the second image or even the wreck in the fourth image.

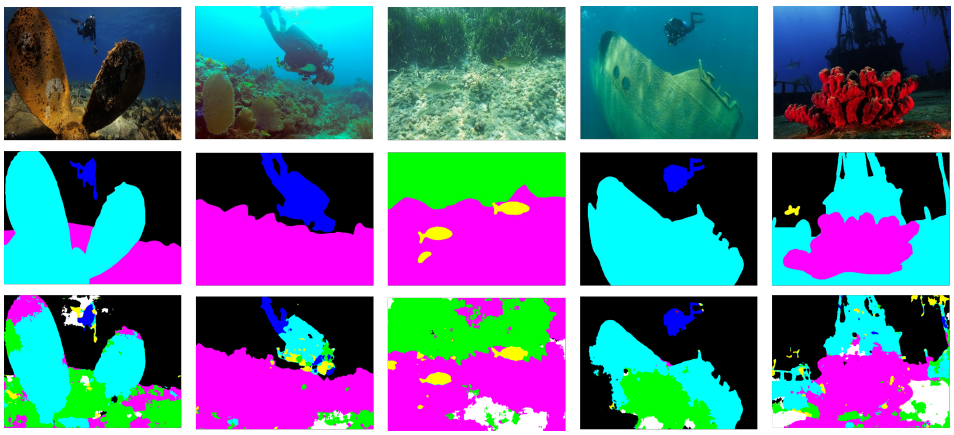

Figure 3. The segmented images obtained with the original dataset for a total of 100 evaluations after every 500 training images are based on the best model, with each different colour representing a different class.

Although the results obtained are not bad, it is important to investigate whether there is a way to increase these numbers to avoid possible errors. Since the investigation of the parameters of the model has already been carried out, the next steps will be to investigate other strategies that can be applied to the training dataset. As mentioned earlier, 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 presents several problems, such as light absorption, turbidity and limited range, which reduces the number of good-quality images. The existing datasets for semantic segmentation training are therefore intended for out-of-water applications. When they are intended for underwater applications, they are sparse in terms of classes (often, a dataset contains only one class, e.g., for fish classification) and has a smaller number of samples (many of them contain 1000/2000 images in total, which is considered a small dataset). In this particular case, the dataset contains a total of 1525 images for training with 8 different classes, which may mean fewer images per class. Also, this dataset is unbalanced, as the class with the highest and lowest representativeness is present in more than 80% and less than 10% of the images, respectively. Another important aspect of this dataset for

semantic segmentation is the need for a suitable annotation for the training process. This task is challenging, as it often has to be completed by the user or with the help of computer tools, but this is very time consuming. The dataset attempts to represent each class under different viewing angles, lighting conditions, sizes, etc., as in the case of the fish class, but in some other cases, this is difficult. Some of the images have poorer visual conditions, with poor boundaries and without correct information about the colour of the objects. In others, there are some inconsistencies in the annotation mask, as it is a difficult and repetitive task. Figure 4 shows some of the problems of the SUIM dataset.

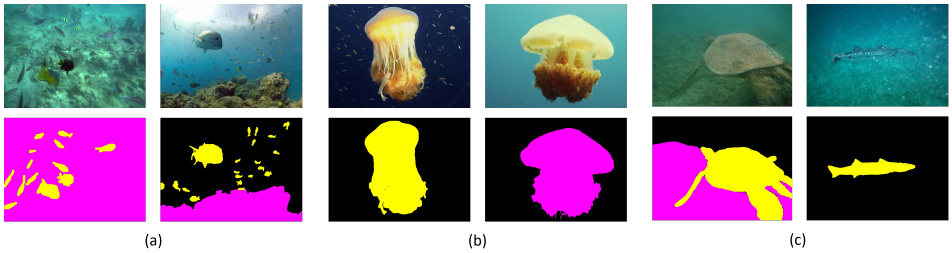

Figure 4. Problems of the dataset: more elements of a class in the original image than in the mask (**a**), some labeled errors when referring to the same class (**b**) and poor quality of the images obtained underwater (**c**). The 3 colours that appear in the second row indicate the 3 classes: Fish (yellow), Reefs (pink) and Waterbody (black).

In this way, two examples can be observed in which there are more fish in the original images than in the ground truth masks (Figure 4a), which can be explained by the fact that labelling is a hard task and is often carried out by hand. Another example is the same class with different labels in the masks (Figure 4b), which can confuse the learning process. Finally, the third example shows the difficulties often encountered with regard to perceiving the boundaries of the objects (Figure 4c). Therefore, it is necessary to review the weak points of the dataset, improve some problems and solve others to improve the final performance (e.g., the poor segmentation).

3. Auxiliary Techniques for Semantic Segmentation

According to the literature and all 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 that can be used for different enhancements. One of them deals with the restoration of true colours, another with filtering techniques to de-noise the image, and others allow the removal of backscatter caused by tiny particles in the water that reflect light back to the camera [19]. After some research in the literature and existing knowledge, it was decided to test only a few techniques, most of which enhance contrast and highlight the boundaries of objects (crucial for a good training process) and which have already performed 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) [20], White Balance (WB) [21] and a third technique derived from diving experience that summarises some enhancements suitable for underwater imaging. Thus, in detail:

1. CLAHE is used to increase the contrast of an image by redistributing light across the image. It is one of the methods of histogram equalisation, but is not applied to the whole image, 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 the object boundaries and some important features. The implementation used is based on the

OpenCv library 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 to look natural to humans. This is not an easy task, as the water absorbs and scatters light differently than in the air and many colours disappear in the depths, where everything looks blue or green. In this context, this technique is used for post-processing, but it is not trivial to achieve the right white balance. There are several variations that occur when implementing this technique, and the approach used [22] shows that realistic degradation of the colour can be accurately minimised (without producing unrealistic colours, such as the purple of the grass, which can affect the results of semantic segmentation). However, the approach used was not tested in underwater environments, which is why some images were produced without quality in the experiments. Thus, different parameter combinations were 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 [23], 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 has been created with the new red colour, it is normalised based on the low and high thresholds, and then the colour matrix is created based on the new values. In summary, this process makes the image appear more natural and with greater contrast.

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 relation to the ground truth, and for this reason, the original masks were used. As can be seen in Figure 5, the results of the enhancement methods vary in comparison to each other and it is necessary to check whether this preprocessing improves the final result. As shown, the results vary and lead to more visually appealing images with some enhancements, while the same enhancement can lead to undesirable images with others.

Figure 5. 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, some Image Quality Assessment (IQA) methods have been developed over time to quantify the perceived image quality [24,25]. They can therefore be used in an underwater context to automatically recognise which enhancement method is the best, for example. There are many IQA methods, but many of them require a

reference image (the ideal image in cases where the goal is to measure the increase in resolution of low-resolution images) and are therefore based on the comparison of two images. In this underwater context, there are no reference images, and for this reason, these methods are discarded to draw conclusions. Therefore, some traditional blind IQA methods, i.e., methods that do not require a reference image, are used to simulate human visual perception and provide quality assessments, such as:

- Blind/No-Reference Image Quality Evaluator (BRISQUE) [26], which is a spatial quality evaluator that finds and evaluates distortions. It deals with local contrast and luminance.
- Perceptual Image Quality Evaluator (PIQE) [27], which combines a set of statistical, structural and perceptual features tuned to human perception.
- Natural Image Quality Evaluator (NIQE) [28], which evaluates natural images based on structural properties of the images such as luminance, contrast and textures.
- Metric presented in [29], which bases its evaluation of image quality on three types of low-level statistical features in both the spatial and frequency domains. Therefore, it quantifies the artefacts of an image and proves to be effective and efficient. During 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 are expected to be.

Another reason for the poor results of semantic segmentation could be that the dataset is often too small to meet the necessary requirements, and obtaining labelled datasets is expensive and time-consuming. Therefore, it is challenging 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 in the input, using smaller datasets for some of the necessary applications, and avoiding overfitting because the training data are 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 imagery presents many challenges, and the model has to cope with all the changes and variations. Therefore, it is crucial in this context to take into account lighting changes, distortions and perspective changes. 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 6 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 obtain more images with only 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 seaweed appears at the bottom of the image. As can be seen in the Figure 7, vertical flipping is useful in some cases, especially when the camera captures images in an upper plane with respect to the object, i.e., images from the bottom of the underwater floor. But it is wrong in cases where the image is taken in front of the object, as can be seen from the air bubbles that can never come down. In many cases, extreme transformations (flipping or rotating) must be avoided.

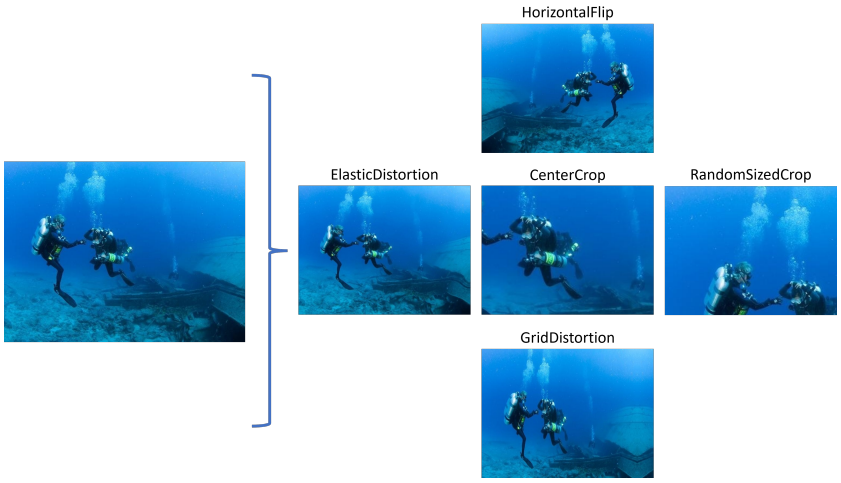

Figure 6. Five examples of possible changes to increase the size of the dataset using a single image.

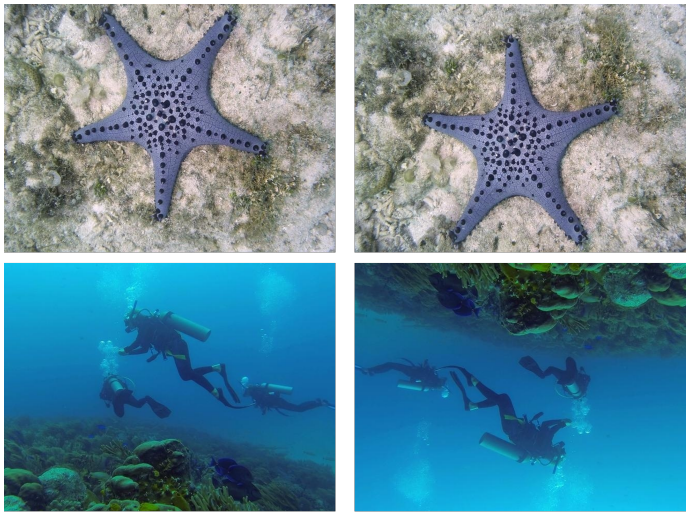

Figure 7. Result from two original images (**a**): in some cases, good variations result, but at the same time images arise that do not fit into the intended context (**b**).

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 are taken from a vehicle that does not rotate much along the x-axis. It is critical that the consistency of the data provided is maintained. The changes must represent reality and the greatest possible number of variations in the real world. There are already many methods of data augmentation that can be implemented to increase the size and diversity of datasets. However, when a segmented dataset is needed, the transformation must be applied to each image while keeping the ground truth consistent. Sometimes it is non-trivial to maintain this so as not to add false information or, for example, 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. But not all transformations require an intervention in the mask, such as the changes in the colour information of the image, where there are no changes in the position or size of the objects. The different transformations can be divided into categories, such as:

- Spatial transformations—these refer to changes in the image coordinate system that affect the position and orientation of objects and the overall image.

- Intensity transformations—these change the appearance of objects, but not their shape, and act at the pixel level.
- Elastic deformations—these 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—this 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.

There are several implementations of these transformations and libraries in the literature for the intended context, such as: ImageDataGenerator by TensorFlow or the transformations by 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 rotating an image and the appearance of black corners. Another option is Albumentations [30], 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 the more specialised, 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, ensuring 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 wide. For this reason, this library was used for this work. In Figure 8, you can see some transformations of the image and ground truth made by this library, e.g., by flipping, cropping and downscaling transformations. By flipping and cropping data, one 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 capture systems, a downscale transformation can be used, as in this context some of these systems provide images with lower resolution. 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.

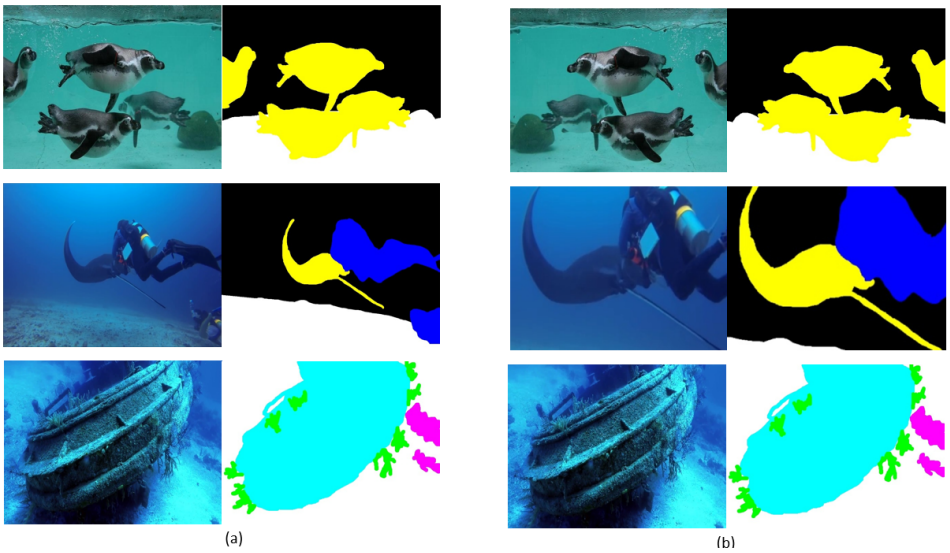

Figure 8. Some examples of original images (**a**) and corresponding transformations (**b**) are shown. The first and second rows represent transformations by flipping and cropping, which also change the mask, and the third row represents a quality degradation where the ground truth does not need to be changed. The second columns represent the ground truth, with each colour representing a different class.

4. Experimental Results

In order to be able to verify the statements made in the literature about the selected procedures 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 explained earlier. All experiments were conducted using an Intel(R)Core(TM) i9-9880 CPU @ 2.3 GHz with 32 GB RAM computer.

4.1. Enhancement Techniques Comparison

As already described, three implementations to enhance the visual quality of the images are tested and compared: CLAHE, white balance and dive correction. It is expected that better visual quality of the images will lead to better training and, consequently, to better semantic segmentation results. As mentioned before, 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 in which the images were acquired, it is important to have an idea of whether these newly generated images lead to a more appealing result or not. Since 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 was created to test them with four previously selected metrics: BRISQUE, PIQE, NIQUE 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 9, where there are images with a quality score of more than 40 (which represents fair quality) and others with less than 20, i.e., excellent quality, according with the methods BRISQUE and PIQE.

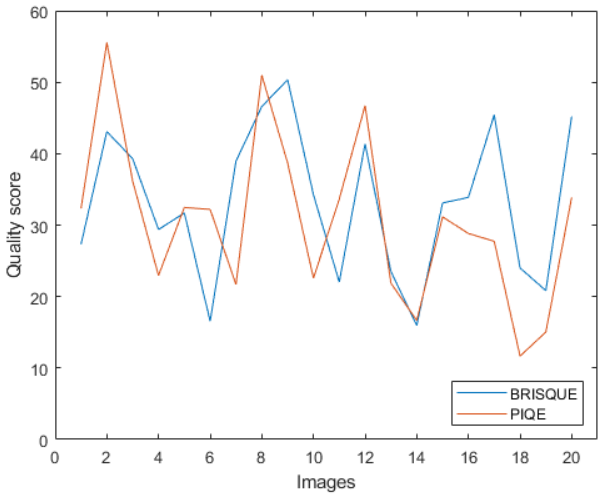

Figure 9. Quality assessment with BRISQUE and PIQE for 20 original images selected to evaluate the quality of images processed with enhancement techniques.

The purpose of this type of test is to ensure that the most appropriate method not only improves the images with low visual quality, but also does not degrade the images that are already of good quality. 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 2nd and 12th images tested are of poor quality and that this is true in terms of human perception (Figure 10a). Furthermore, both metrics indicate that the 14th and 18th images are of high quality, but this conclusion is not clear in terms of human perception, as shown in Figure 10b.

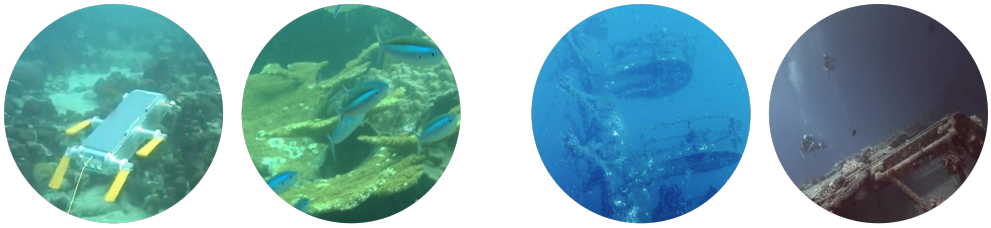

Figure 10. Two examples of bad visual images (**a**) and good visual images (**b**) of the original set, according with the methods BRISQUE and PIQE.

To see if it is possible to determine which method of enhancement technique is most appropriate in the underwater context according to these assessment methods, the quality tests are performed for the three methods of enhancement and with the four selected metrics. The results obtained are presented in Table 2. The obtained values of the metrics that exceed the results obtained with the original set are highlighted in bold.

Table 2. Comparative results of the mean of the quality score of the Dive, WB and CLAHE corrections for a total of 20 images, using PIQE, 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	PIQE	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**

It is also important to point out that the BRISQUE and the PIQE represent the results on a scale of 0 to 100, while the NIQE and the sr_metric are based on a scale of 0–10. Furthermore, for the sr_metric, the higher the score, the higher the quality of the image. For the other metrics, the relationship is reversed, i.e., the best results are achieved with a lower quality score. As you can see, with the first metric, the overall quality of the results obtained decreases as the quality score increases. With BRISQUE, the results of 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, wich 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 13th image, the score for the quality of the image obtained increases sharply (by more than 18 units) compared to the original visual quality for the BRISQUE metric. This therefore 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 obtained for PIQE and NIQE shows a decrease of 6 and 0.3, respectively, indicating a deterioration in quality. With the sr_metric, the results obtained do not differ from each other. In this way, it is not possible to say with absolute certainty that one method is better than the other at 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 offers the best results for segmentation. The three methods are thus applied to the whole training dataset, and also to the test images. For this test, four different datasets were created and the model was trained with a total of 50,000 images, with evaluations after 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 3.

Table 3. Comparative results of best measured model performance [%] 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.

Method	Overall Acc	Mean Acc	Mean IoU	IoU 0	IoU 1	IoU 2	IoU 3	IoU 4	IoU 5	IoU 6	IoU 7
Original	79.7	64.4	53.1	87.2	62.7	29.5	58.8	16.6	51.6	59.1	69.6
CLAHE	80.4	64.0	**53.3**	87.0	61.6	25.7	**59.1**	14.1	**54.2**	**60.5**	64.2
WB	80.0	62.5	51.9	86.6	58.2	**30.0**	**59.7**	8.5	**55.8**	56.8	59.4
Dive	**80.5**	63.4	52.3	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, as it is calculated with respect to all classes. However, since we have eight 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 (greater than 3%). For this reason, this enhancement method seems to be the best, but the visual results should be observed. Figure 11 shows the results obtained by applying the three enhancement methods to the images. Looking at the visual results, it is possible that the CLAHE results are generally better because it is the best method for observing the body of the human diver and the coral 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 because it is a difficult class to model. The other implementation stands out compared to the CLAHE method in the robot class, but it is not significant because it is only one of eight classes and does not provide good segmentation results. In this way, the best result is obtained using the CLAHE method, which slightly improves the final result and has a higher overall accuracy, i.e., a higher rate of well-detected pixels. This result was to be expected, as the contrast and thus the visibility of the main features of the images 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 50,000) after 90 evaluations, i.e., using a total of 45,000 training images instead of 50,000 for white balance enhancement or 47,500 for dive correction. For this reason, the CLAHE method is chosen as the best and most appropriate for the context and for use in the next set of experiments, if necessary.

Figure 12 shows an example of the improvement in the final segmentation results using the CLAHE method in the fish and wreck classes compared to the original results.

4.2. Data Augmentation Influence

As far as data augmentation 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 be larger than necessary and the processing time would increase greatly without necessity. To create the datasets needed for the following tests with data augmentation, a total of eight transformations were used, which can be seen in Figure 13. 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 influence of data augmentation was tested in an isolated way, i.e., without considering visual enhancement methods.

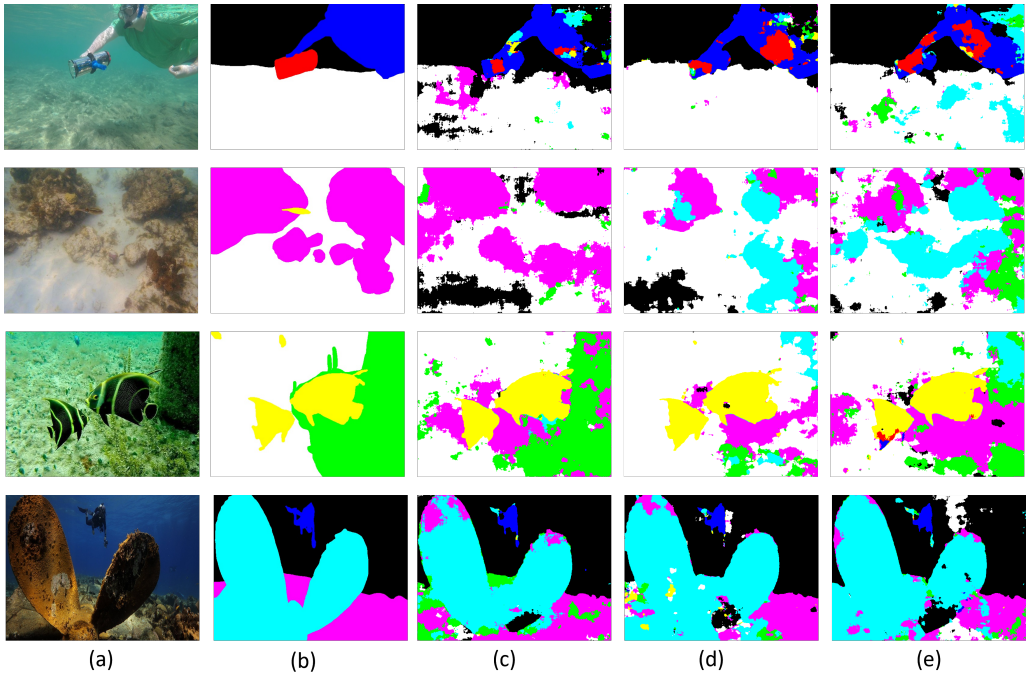

Figure 11. Qualitative comparison results of 4 different examples (**a**,**b**): the CLAHE method (**c**), WB method (**d**) and dive correction (**e**) using the best model in a total of 100 validations, each after 500 training images. The different colours represent the different classes along the results.

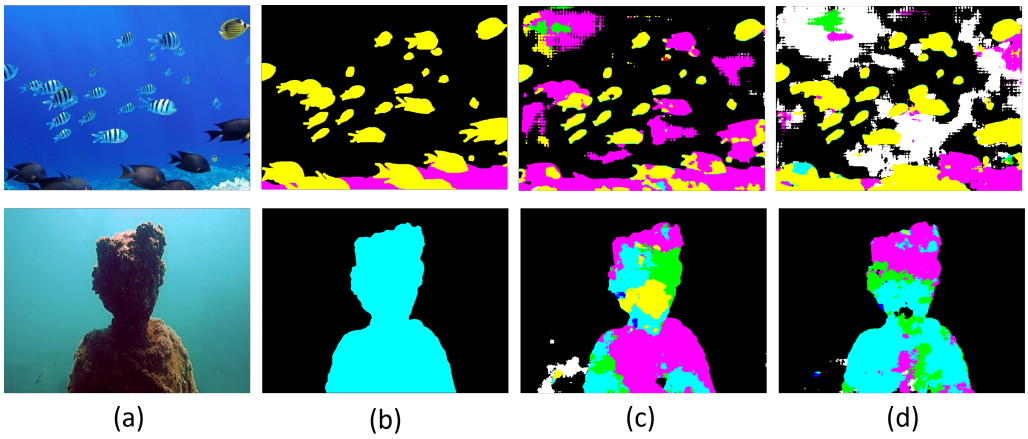

Figure 12. Comparison of the qualitative results of 2 different examples (**a**,**b**), between the result obtained with the original set (**c**) and the CLAHE method (**d**) using the best model for a total of 100 validations. The different colours represent the different classes along the results.

4.2.1. General Data Augmentation

The first experiment with data augmentation was conducted on a dataset in which five of the augmentations were applied to each image in the training dataset. This provided a total number of 9144 images for the training to be able to evaluate the effects of this phase in a larger dataset, as 1500 images is a small dataset for the intended context. Figure 14 illustrates the distribution of classes across images. It can be seen that the number of samples per class increases and the distribution of classes remains the same.

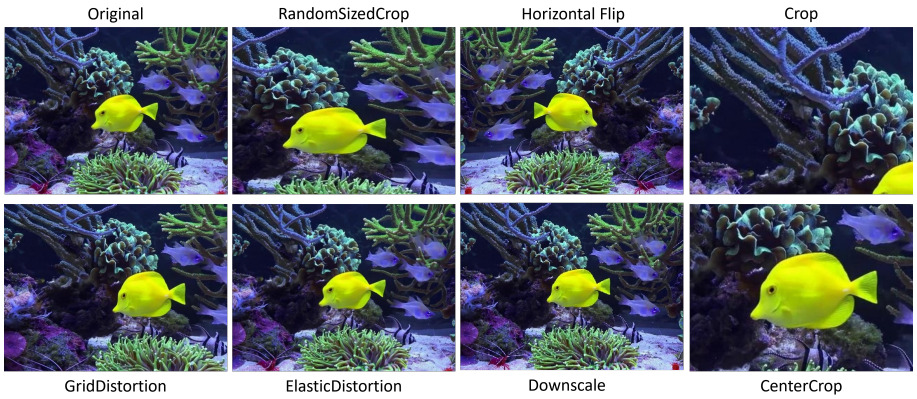

Figure 13. Eight transformations used in the creation of the augmented datasets for training.

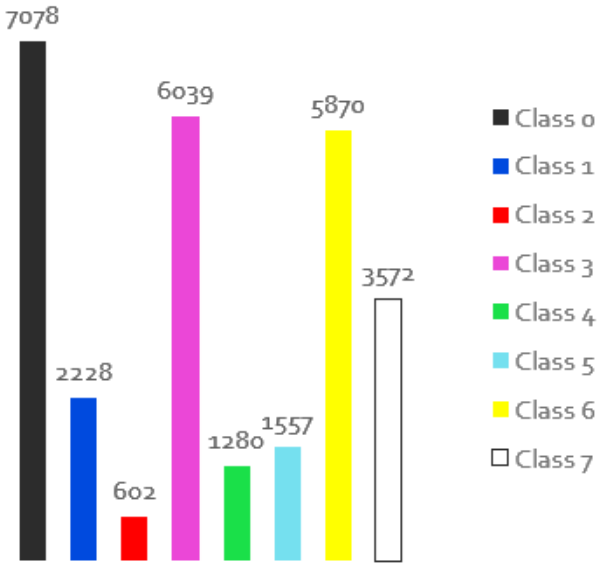

Figure 14. Distribution of the individual classes for the images of the training set with data augmentation.

To allow a fair comparison and because the new dataset is larger, the original dataset was run for 50,000 training images, with the evaluations carried out for 1000 images each (about two-thirds of the total number of images). Therefore, the new dataset with data augmentation (DA) was evaluated after every 6000 images for a total number of 300,000 iterations, with the aim of having the same number of evaluations at the end, i.e., 50 evaluations. Table 4 summarises the result for the best iteration for the original dataset and the dataset with data augmentation (iteration 41 and iteration 47, respectively). Looking at the results, we can see that increasing the number and variety of samples (under the same conditions for the train) increases the general value of the mean IoU by about 2.5% (bold type on the table), which is a relevant result because there are five classes whose IoU is increased. There are two classes in which the behavioural performance decreases, but one of them is a difficult class to model (plants) and another is a general class because it includes rocks. The value of this class is still useful in the intended context.

Table 4. Performance measures [%] for the best result obtained with the original dataset and the dataset with data augmentation 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.

	Overall Acc	Mean Acc	Mean IoU	IoU 0	IoU 1	IoU 2	IoU 3	IoU 4	IoU 5	IoU 6	IoU 7
Original	80.6	62.9	51.8	87.6	60.9	20.9	60.2	11.6	54.1	55.5	64.0
DA	80.4	**65.0**	**54.1**	85.7	**64.2**	**38.6**	**61.3**	8.6	**54.3**	**61.1**	58.8

Figure 15 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 the data segmentation learns faster.

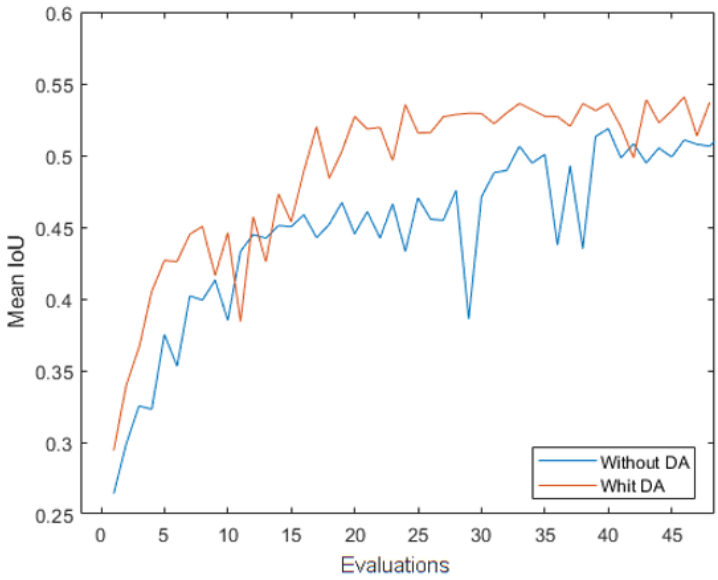

Figure 15. Mean IoU measurement along the 50 evaluations with the original dataset and the data augmentation dataset.

Figure 16 shows the same analysis, but in relation to the individual classes of the process. It can be stated 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. With respect to the worst classes (plants and robots), it is important to point out that the results start earlier with data augmentation, which means 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, which means that these classes are less represented than the others.

Figure 17 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 identified as robots occur in large numbers, and wrecks and human divers are better delineated. Plants and robots are the classes that provide the worst results, as they remain the classes with low representativeness.

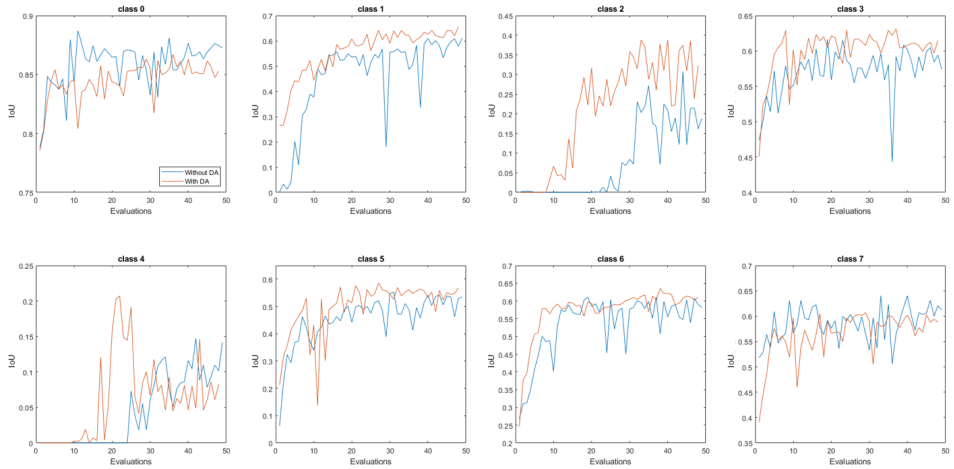

Figure 16. IoU measurement along the 50 evaluations with the original dataset and the data augmentation dataset, for each class.

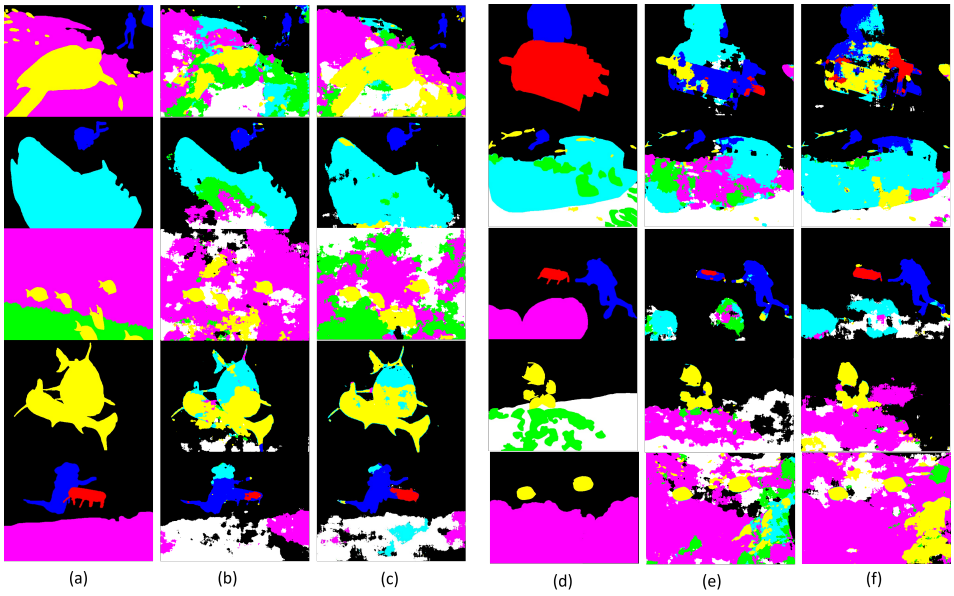

Figure 17. Qualitative comparison of the results of 10 different examples of semantic segmentation (**a**,**d**) using the original dataset (**b**,**e**) and the data augmentation set (**c**,**f**) for the training process, in which the best model was selected in a total of 50 validations. The different colours represent the different predicted classes along the results.

4.2.2. Data Augmentation of the Less Representative Classes

It is clear that segmentation is good for improving the results because it produces 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. If we look at the distribution of the original dataset used for training (see Figure 2), 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 five transformations (CenterCrop, Crop on the left, Grid and Elastic Distortion and Dowscale) were applied to the less representative classes. The generated dataset includes a total of 7312 images. This is a larger dataset, but it still needs to be verified if this number of images is sufficient to model the eight classes. The new distribution of classes in all images is shown in Figure 18. 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 in the worst classes is not proportional (while the robot class increases by six times, the wrecks increase by seven 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, may affect this result.

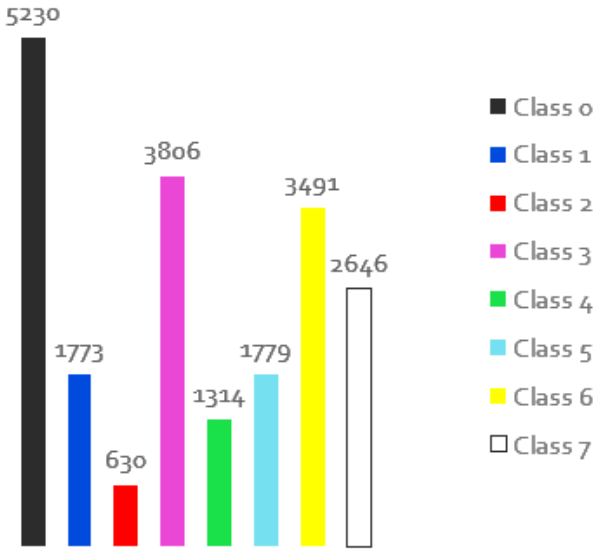

Figure 18. Distribution of the individual classes for the images of the training set when applying the data augmentation in different numbers according to the class.

As in the previous experiment, training was performed on the images in this new dataset with data augmentation (DA1) after every 5000 images (about two-thirds of the total number of images in this dataset) for a total number of 250,000 iterations, with the aim of obtaining the same number of evaluations at the end as the original result presented before. Table 5 summarises 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 5. Performance measures [%] for the best result obtained with the original dataset and the new dataset with data augmentation (DA1) 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.

	Overall Acc	Mean Acc	Mean IoU	IoU 0	IoU 1	IoU 2	IoU 3	IoU 4	IoU 5	IoU 6	IoU 7
Original	80.6	62.9	51.8	87.6	60.9	20.9	60.2	11.6	54.1	55.5	64.0
DA1	80.2	**66.0**	**54.4**	86.3	**61.1**	**37.7**	**61.2**	**23.5**	53.4	54.8	57.9

As shown, both the value of the mean IoU and the mean accuracy increase because there are some classes that increase both the accuracy in detecting the pixels and the IoU. The classes that are more 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. However, the plants and the robots increase the value of IoU by 12% and 16%, respectively, which is good because 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 because in most cases, the robots appear on the image together with the human divers. 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. Curiously, this model was chosen because it has the best value for the mean IoU during the 250,000 training iterations, but

this best value corresponds to iteration number 22 (fewer evaluations). Therefore, Figure 19 shows the segmentation obtained for images with the classes augmented in the training process: Wrecks, Plants and Robots. For these particular 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.

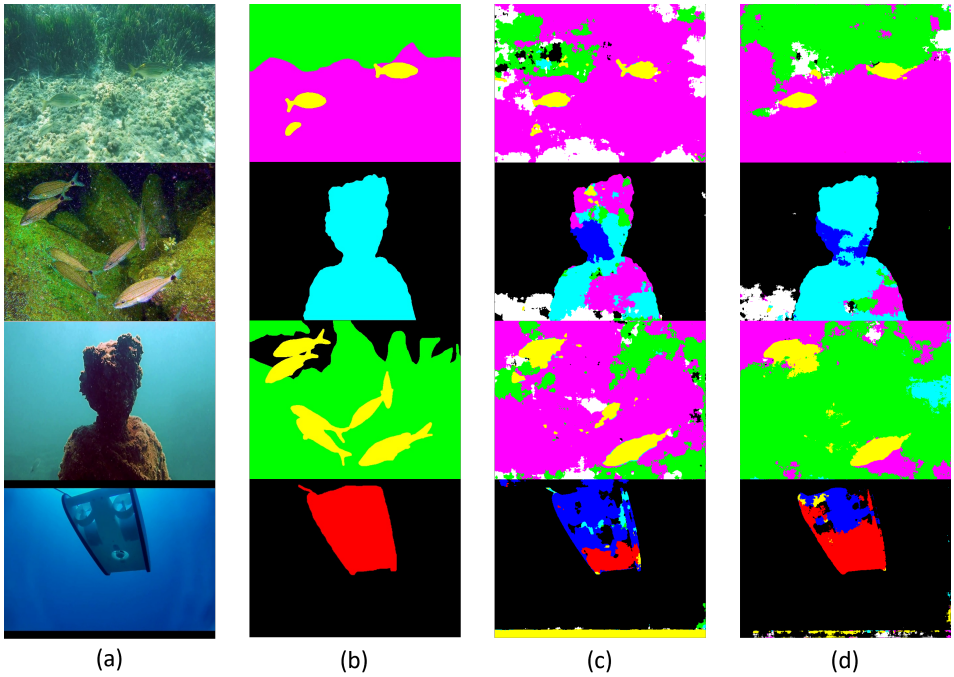

Figure 19. Qualitative comparison results of 4 different examples of semantic segmentation (**a**,**b**), obtained from the original training set (**c**) and the data augmentation set (**d**), using the best model in a total of 50 validations for each case. The different colours represent the different predicted classes along the results.

4.2.3. Data Augmentation of the Isolate Classes

The last result, related to data augmentation, has the main objective of testing if it is possible to increase the performance of an isolated class via 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 seven semantic segmentations of those described previously. The new dataset shows the distribution of the graph of Figure 20, obtaining a total of 2252 images.

The number of images with the robot class increases strongly and its representation reaches more than 25%, which is a good result for a dataset with eight 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 often co-occur with the robot. Therefore, this dataset was trained with a total of 75,000 images, evaluating all 1500 images to perform a total of 50 evaluations, and the best model was determined after 36 evaluations. The result of this best evaluation provided 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 increased 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 21 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 was always lower than when the images with data augmentation were used, the values in the dataset with data augmentation began to show. In the original case, the class only shows results higher than zero (also unusable) after 25 evaluations.

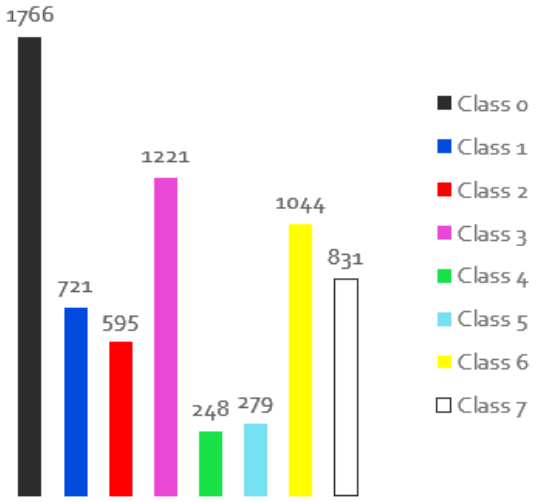

Figure 20. Distribution of the individual classes for the images of the training set when applying the data augmentation only in the robot class.

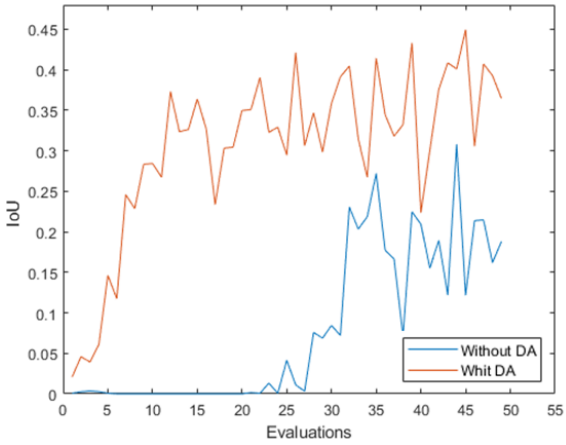

Figure 21. IoU measurement for the robot class (class 2) along the 50 evaluations with the original dataset and the data augmentation dataset.

Figure 22 shows the result of some images with robot class present, for which the best model obtained in this test was used.

Figure 22. 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 different colours represent the different predicted classes along the results.

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 augmented with seven data augmentations. The difference is that only about 100 of the total number of plants were augmented because this class occurs in many images with very small sizes or only in one corner of the image and the transformation produces images without this class (see

Figure 23). The remaining images were retained in the dataset without data augmentation in a similar way as all images without plant class. In this way, we obtained 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 24. Therefore, this dataset was trained with a total of 75,000 iterations, evaluating all 1500 images to perform a total of 50 evaluations, and the best model was obtained after 45 evaluations. As for the result, similar behaviour as for the robot class was observed, i.e., it started with values above 0% and obtained higher results during the 50 evaluations (see Figure 25).

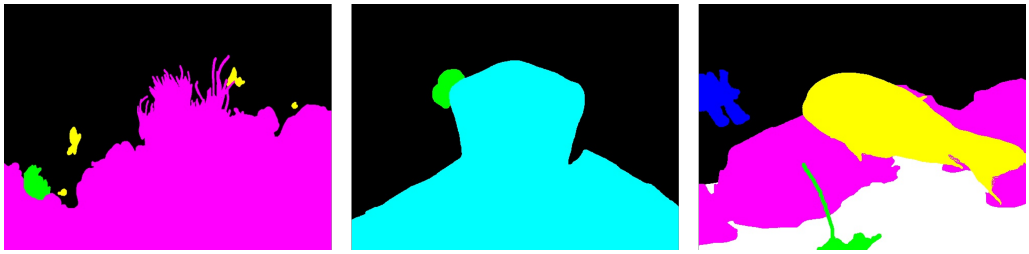

Figure 23. Demonstration of the size and location on the original set of the plants in some of the images. The different colours represent the different classes in the original masks.

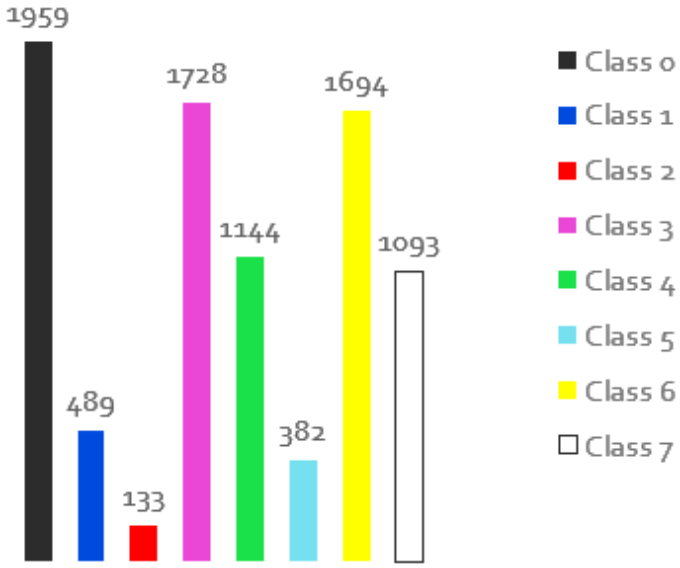

Figure 24. Distribution of the individual classes for the images of the training set when applying the data augmentation only in the plant class.

In this case, the best value of the mean IoU (53%) does not refer to the best value of the IoU for the plants, as all classes are considered in this metric. The value obtained for the IoU for the plant was 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.

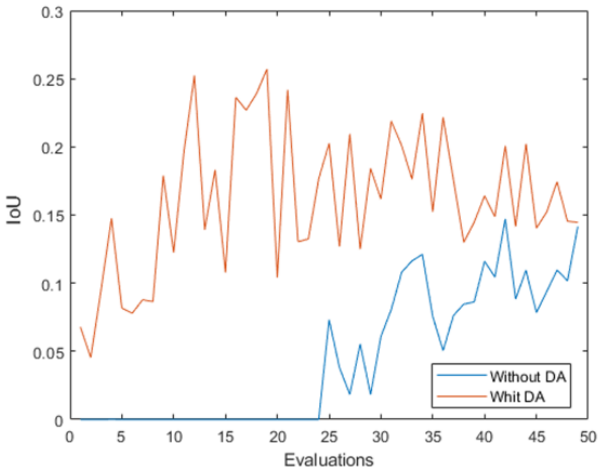

Figure 25. IoU measurement for the plant class (class 4) along the 50 evaluations with the original dataset and the data augmentation dataset.

4.3. Data Augmentation with CLAHE Method

In order to evaluate the effect of using both strategies simultaneously and to see if the results improved, we decided to use the DA1 dataset, i.e., the more balanced dataset of the previous tests and the better enhancement technique (CLAHE). Therefore, training with the images from this new dataset (DA _C)) was performed after every 5000 images for a total number of 250,000 iterations to perform 50 evaluations. Table 6 shows the comparison between the new results and the results obtained in the test with the data augmentation only. In bold type, you can see whether the results achieved with both strategies outperform the result with only the data augmentation. This allows you to clarify whether it makes sense to use both strategies or whether the data augmentation is sufficient.

Table 6. Performance metrics [%] comparison for the best result with the more balanced dataset with data augmentation (DA) and this dataset with CLAHE method (DA_C) applied for a total of 50 validations. For each result, you can see whether using both strategies is better (in bold) than using only the data augmentation.

	Overall Acc	Mean Acc	Mean IoU	IoU 0	IoU 1	IoU 2	IoU 3	IoU 4	IoU 5	IoU 6	IoU 7
DA1	80.2	66.0	54.4	86.3	61.1	37.7	61.2	23.5	53.4	54.8	57.9
DA_C	**81.8**	**67.6**	**56.0**	**87.9**	**66.7**	**39.4**	**62.4**	12.8	**53.9**	**61.0**	**63.9**

From the results, we can see that all general performance metrics are increasing, especially overall accuracy, which has been the best metric since the first test. Despite a lower result, the mean accuracy improves by 1.5% and in terms of mean IoU, a value of 56% is achieved, which is the best result ever obtained with this dataset (a good result, given the number of evaluations that are performed). If we look at the results of the individual classes, 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 class of plants 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 was shown that using both strategies together is a good plan to improve the results in general. Figure 26 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 fishes, represent the complete object).

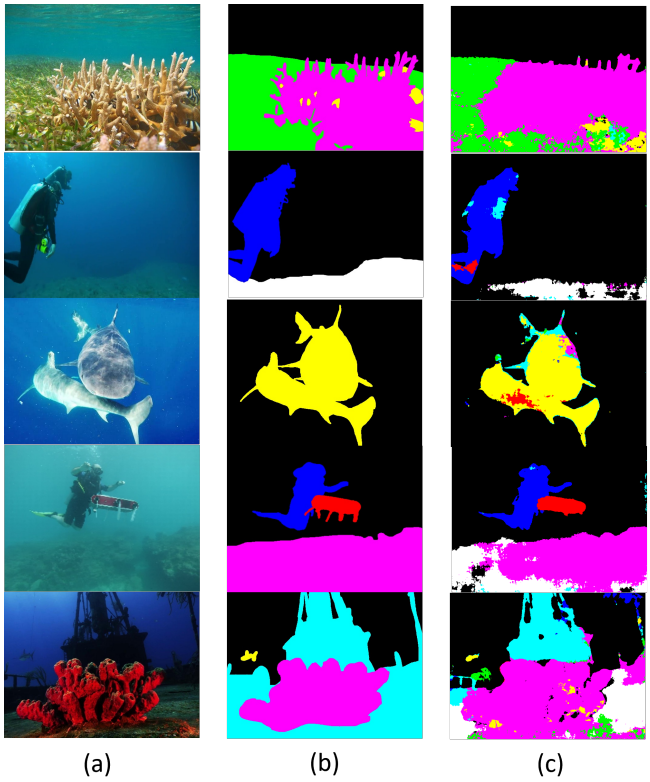

Figure 26. Results of 5 different examples of semantic segmentation (**a**,**b**), obtained from the use of both strategies at the same time: data augmentation of the less representative classes and CLAHE method (**c**). The different colours represent the different classes in the original masks and predicted results.

5. Conclusions

The main objective of this work was to evaluate whether there are approaches commonly used in outdoor environments that can be applied in visual segmentation methods for the underwater context. If so, how much they could improve the overall performance of this task? We wanted to find out how to build a more balanced dataset without requiring data collection, which is already challenging and expensive in terms of time and cost, in order to extend these methods to other domains and application tasks. For this purpose, we start from the results of a previous study conducted by the authors. In that work, four literature implementations for semantic segmented applications were tested and Segnet stood out for its performance, which is why the same method is also used in this work. The lack of good datasets with which to test these applications is still a problem, namely datasets for semantic segmentation, which also requires the labelled ground truth for each image. Therefore, the SUIM dataset was chosen because, although it contains only 1500 images for training, it presents different perspectives, sizes and conditions of eight classes that are commonly encountered in this world. Although not all objects intended for our applications have been covered yet and there are some errors in the labelled masks, the authors resort to this dataset in both papers to perform the evaluations. Using the Segnet in a training with 50,000 training images and evaluations after every 500 images gives an overall accuracy of 80%, a mean accuracy of 64% and a mean IoU of 52% for the best result in 100 evaluations, which is our starting point. The model shows low performance, with only three of the total classes achieving more than 60% and two classes achieving less than 30%, which could be dangerous in a real mission. Therefore, some strategies need to be tested to see if it is possible to improve these results. For this purpose, approaches to improve image quality and data augmentation were tested. Therefore, the main conclusions about the results are as follows:

- The better the quality of the visual images, the better the segmentation results are expected to be. Therefore, in the first experiments, an attempt was made to determine which enhancement method is the best using some IQA metrics that do not require a reference image. Using 20 previously selected heterogeneous images, it was found (looking at the average quality scores) that CLAHE and dive corrections improved the results for three out of four metrics. However, when looking at individual cases, some inconsistencies were found, so it cannot be said with certainty which enhancement method is best suited to achieve better segmentation results.
- The first test of segmentation results was carried out with 50,000 training images. The evaluation was performed after 500 images were randomly selected, i.e., after 100 evaluations, with the main aim of finding out which enhancement techniques can improve segmentation performance. The test uses 100 images which have been selected beforehand and are the same for all tests. Looking at the quantitative results, we can see that the CLAHE method is the only one that increases the mean IoU value, but with a small increase. As for 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.

The next tests have the main objective of testing the impact of data augmentation for the intended application, since the dataset used is small and eight transformations were used to create the new datasets used in the tests.

- In the first test of data augmentation, a new dataset with a total of about 9000 images is created using five random data augmentations. The dataset is used for training with a total of 300,000 training images, which are evaluated after every 6000 images (two-thirds of the total dataset). A new test of the original images is performed under the same conditions (a total of 50,000 training iterations and evaluations after every 1000 images to obtain a total of 50 evaluations). 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 some classes with more diversification to be used in the learning process. There are five 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 in which the increase in the number of images also leads to an unbalanced dataset. The classes that decrease their performance decrease by values that are still good and useful in a real context, i.e., they are not dangerous.
- The following test is to show what happens when we augment the classes with a lower representativeness in the images. For this purpose, the images with plants, robots and wrecks are augmented with seven transformations and the rest with only two. However, it is impossible to isolate classes, and sometimes when we augment a class with an image, we augment more than this class, as there is more than one class in the original. The result is a slightly more balanced dataset in which the higher representative class occurs in about 70% instead of 80% of the images and the three less representative classes occur on average in about 17% instead of 12%. Under the same conditions as the previous test, the performance for the mean IoU value is 54.4%, with the robots improving their result and the plants having an IoU value of 23.5% for the best model obtained during training.
- To test whether it is possible to increase the value of the isolated classes, two tests are performed: one with the robot class that is least representative in the original dataset, and another with the plants; this is shown to be the most difficult class to learn. To achieve this, the images in which the robot class occurs are isolated and seven transformations are performed so that the robot is no longer the least representative class. The IoU value for this class increases to a value of 41.4% (a value that has

never been reached for this class before). When examining the plants, it is not only necessary to isolate the images of the plant class, but also to select only some of them for data augmentation. In many images, the class appears with few pixels and the transformations are images without the class. If we compare the results of the class with this data augmentation and the original case, we can see that the class has higher values than 0% in the data augmentation test at first. 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. These tests show that it is possible to increase the values of a class by feeding the model with a variety of images.

- The last test has the main objective of seeing if the technique of data augmentation and enhancement provides good results. For this purpose, the CLAHE method was applied to the data augmentation of the less representative classes (DA1). The results obtained are very good, with a mean IoU value of 56% and an increased value of overall accuracy of 81.8% (the highest value at this time). The mean accuracy value is also increasing, which means that the results are better in different classes. This is not the case for plants, which may mean that the enhancement may confuse the model in the regions of this class (e.g., set 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 frequently.

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 five classes with more than 60% and only one with less than 30% for the IoU value. This shows that the results are more reliable and can be more easily used in real contexts. It is important to mention that the results are different from the original results, which were obtained with more training. This is a good result because it is to be expected that with longer training periods, the values of some classes will achieve higher performance, as they are more difficult to train. Both strategies are thus 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 plans include automating the data augmentation process to apply more transformations without worrying about the associated acquisition, e.g., rotational or flipping transformations (which are allowed in some cases and avoided in others, when it is not practical), and testing performance with images acquired during real missions of the robots. In addition, both strategies to improve segmentation results will be tested using the current methods to see if their application will always improve the final results. It is also important to include other objects in the dataset that are important for the movement of the vehicles, e.g., pipelines, anchors, etc.

Author Contributions: Conceptualization, A.N. and A.M.; Methodology, A.N.; Validation, A.N.; Writing (original draft preparation), A.N.; Writing review and editing), A.N. and A.M.; Supervision, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research is financed by FCT—Fundação para a Ciência e a Tecnologia—and by FSE—Fundo Social Europeu through of the Norte 2020—Programa Operacional Regional do Norte—through of the doctoral scholarship SFRH/BD/146461/2019.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data available on request.

Acknowledgments: The authors would like to acknowledge Minnesota Interactive Robotics and Vision Laboratory for providing the SUIM Dataset [11] that supports this research.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BRISQUE	Blind/No-Reference Image Quality Evaluator.
CLAHE	Contrast Limited Adaptive Histogram Equalization.
DA	Data Augmentation.
FCNN	Fully Convolutional Networks.
IQA	Image Quality Assessment.
IoU	Intersection-over-Union.
NIQE	Natural Image Quality Evaluator.
PIQE	Perceptual Image Quality Evaluator.
PSPNet	Pyramid Scene Parsing Network.
SUIM	Segmentation of Underwater IMagery.
WB	White Balance.

References

1. Chen, W.; He, C.; Ji, C.; Zhang, M.; Chen, S. An improved K-means algorithm for underwater image background segmentation. *Multimed. Tools Appl.* **2021**, *80*, 21059–21083. [CrossRef]
2. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Rodríguez, J.G. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.
3. Zhou, Y.; Wang, J.; Li, B.; Meng, Q.; Rocco, E.; Saiani, A. Underwater Scene Segmentation by Deep Neural Network. In Proceedings of the 2nd UK-RAS Robotics and Autonomous Systems Conference, Loughborough, UK, 24–31 January 2019; pp. 44–47. [CrossRef]
4. Drews, P., Jr.; Souza, I.; Maurell, I.; Protas, E.; Botelho, S. Underwater image segmentation in the wild using deep learning. *J. Braz. Comput. Soc.* **2021**, *27*, 12. [CrossRef]
5. Wang, J.; He, X.; Shao, F.; Lu, G.; Hu, R.; Jiang, Q. Semantic segmentation method of underwater images based on encoderdecoder architecture. *PLoS ONE* **2022**, *17*, e0272666. [CrossRef]
6. Garcia-D'Urso, N.; Galán Cuenca, A.; Pérez-Sánchez, P.; Climent i Pérez, P.; Guilló, A.; Azorin-Lopez, J.; Saval-Calvo, M.; Guillén-Nieto, J.; Soler-Capdepón, G. The DeepFish computer vision dataset for fish instance segmentation, classification, and size estimation. *Sci. Data* **2022**, *9*, 287. [CrossRef]
7. Saleh, A.; Laradji, I.; Konovalov, D.; Bradley, M.; Vázquez, D.; Sheaves, M. A Realistic Fish-Habitat Dataset to Evaluate Algorithms for Underwater Visual Analysis. *Sci. Rep.* **2020**, *10*, 14671. [CrossRef] [PubMed]
8. Beijbom, O.; Edmunds, P.; Kline, D.; Mitchell, B.; Kriegman, D. Automated Annotation of Coral Reef Survey Images. In Proceedings of the CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1170–1177. [CrossRef]
9. Alonso, I.; Yuval, M.; Eyal, G.; Treibitz, T.; Murillo, A.C. CoralSeg: Learning Coral Segmentation from Sparse Annotations. *J. Field Robot.* **2019**, *36*,1456–1577. [CrossRef]
10. Fabbri, C.; Islam, M.J.; Sattar, J. Enhancing Underwater Imagery Using Generative Adversarial Networks. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 7159–7165. [CrossRef]
11. Islam, M.J.; Edge, C.; Xiao, Y.; Luo, P.; Mehtaz, M.; Morse, C.; Enan, S.S.; Sattar, J. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25-29 October 2020; pp. 1769–1776.
12. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
13. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239. [CrossRef]
14. Jonathan, L.; Evan, S.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015.
15. Shah, M.P. Semantic Segmentation Architectures Implemented in PyTorch. 2017. Available online: https://github.com/meetshah1995/pytorch-semseg (accessed on 6 March 2023) .
16. Nunes, A.; Gaspar, A.; Matos, A. Comparative Study of Semantic Segmentation Methods in Harbour Infrastructures. In Proceedings of the OCEANS 2023, Limerick, Ireland, 5–8 June 2023; pp. 1–9. [CrossRef]
17. Alomar, K.; Aysel, H.I.; Cai, X. Data Augmentation in Classification and Segmentation: A Survey and New Strategies. *J. Imaging* **2023**, *9*, 46. [CrossRef] [PubMed]
18. Jian, M.; Liu, X.; Luo, H.; Lu, X.; Yu, H.; Dong, J. Underwater image processing and analysis: A review. *Signal Process. Image Commun.* **2021**, *91*, 116088. [CrossRef]

19. Sanila, K.H.; Balakrishnan, A.A.; Supriya, M.H. Underwater Image Enhancement Using White Balance, USM and CLHE. In Proceedings of the 2019 International Symposium on Ocean Technology (SYMPOL), Ernakulam, India, 11–13 December 2019; pp. 106–116. [CrossRef]
20. Pizer, S.; Johnston, R.; Ericksen, J.; Yankaskas, B.; Muller, K. Contrast-limited adaptive histogram equalization: Speed and effectiveness. In Proceedings of the First Conference on Visualization in Biomedical Computing, Atlanta, GA, USA, 22–25 May 1990; pp. 337–345. [CrossRef]
21. Ramanath, R.; Drew, M.S. White Balance. In *Computer Vision: A Reference Guide*; Springer US: Boston, MA, USA, 2014; pp. 885–888. [CrossRef]
22. Afifi, M.; Brown, M.S. What Else Can Fool Deep Learning? Addressing Color Constancy Errors on Deep Neural Network Performance. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–3 November 2019. [CrossRef]
23. Bornfree, H. Dive and Underwater Image and Video Color Correction. 2022. Available online: https://github.com/bornfree/dive-color-corrector (accessed on 17 July 2023).
24. Reibman, A.R.; Bell, R.M.; Gray, S. Quality assessment for super-resolution image enhancement. In Proceedings of the 2006 International Conference on Image Processing, Atlanta, GA, USA, 8–11 October 2006; pp. 2017–2020. [CrossRef]
25. Zhou, W.; Wang, Z.; Chen, Z. Image super-resolution quality assessment: Structural fidelity versus statistical naturalness. In Proceedings of the 13th International Conference on Quality of Multimedia Experience (QoMEX), Virtual, 14–17 June 2021; pp. 61–64. [CrossRef]
26. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef] [PubMed]
27. N, V.; D, P.; Bh, M.C.; Channappayya, S.S.; Medasani, S.S. Blind image quality evaluation using perception based features. In Proceedings of the 2015 Twenty First National Conference on Communications (NCC), Bombay, Mumbai, 27 February–1 March 2015; pp. 1–6. [CrossRef]
28. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [CrossRef]
29. Ma, C.; Yang, C.Y.; Yang, X.; Yang, M.H. Learning a no-reference quality metric for single-image super-resolution. *Comput. Vis. Image Underst.* **2017**, *158*, 1–16. [CrossRef]
30. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [CrossRef]