Underwater Image Semantic Segmentation with Weighted Average Ensemble

Muhammad Hidayat Jauhari(✉) and Noramiza Hashim

Multimedia University, Cyberjaya, Malaysia
muhidayat4800@gmail.com

Abstract. Underwater image segmentation is a method that could help with underwater exploration because it is useful and impactful in the understanding and study of the marine environment. However, it is a difficult and challenging task compared to regular image segmentation due to the nature of the images themselves, which are of lesser quality, as well as the limited availability of publicly accessible datasets. In this work, several deep learning-based approaches were implemented and a solution for underwater image segmentation was proposed. The proposed method was developed using a smaller dataset with low resolution images. The proposed method consisted of several image segmentation deep learning models such as U-Net, LinkNet, and Feature Pyramid Network (FPN) with different encoders specifically Inception-V3 and ResNet34. The weighted average ensemble method was used to combine the results of each models mentioned to identify the optimized combination. The proposed method was then compared with the individual models to provide comparison and benchmark on the ensemble approach to the single model approach. The proposed method achieved an accuracy of 62.48% where it outperformed all individual models. As a result, aggregating expected results from numerous models gives better performance when compared to individual model predictions.

Keywords: Underwater Image · Weighted Average Ensemble · Deep Learning · Neural Network · Semantic Segmentation

1 Introduction

Due to the vast size of the ocean, which cover around 70% of the Earth, many things have yet to be discovered based on underwater exploration. Through the data and information resulting from underwater explorations, many are useful to scientists, engineers, and researchers. For example, information like how human activities could affect the ocean and how the ocean could be affected by the Earth's environmental, weather and climate changes could be highlighted for discovering new ideas that could be beneficial towards humans and the Earth [1]. Underwater exploration could also give insights for natural disasters such as earthquakes and tsunamis [1].

Scientists, engineers, and researchers have invented a range of methods for carrying out underwater exploration because of its importance. Underwater image processing

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 533–543, 2022.
https://doi.org/10.2991/978-94-6463-094-7_43

is one of the approaches employed which includes segmentation, enhancement, and restoration, where semantic segmentation is the process where objects in a digital image is segmented into subgroups or categories of objects.

One of the methods to perform underwater image semantic segmentation is by using deep learning techniques. The advancement of various deep learning models has helped in yielding better performance such as better accuracy rates [2].

Underwater image segmentation is quite challenging compared to traditional image segmentation because of the underwater environment. Underwater images usually have low visibility due to low contrast, light scattering and noises as well as having a green or blue hue due to the environment itself. Other than that, the limited availability of publicly available dataset also poses a challenge as it affects the performance of underwater image segmentation models.

This paper aims to propose a method that combines multiple deep learning models such as U-Net [3], LinkNet [4], and FPN [5], with different encoders such as ResNet34 [6] and Inception-V3 [7]. The fusion of the result using weighted average ensemble. This study was conducted using a smaller size dataset with low resolution images, which is another difficulty in image segmentation.

2 Background Study

Various techniques have been devised, developed, and refined to further improve underwater picture segmentation and its application as the computer vision, machine learning, and data mining fields have advanced. This section will address existing approaches proposed by the implementation of deep learning models, as well as a brief discussion of weighted average ensembles.

2.1 Underwater Image Segmentation Based on Deep Learning

In [8], the authors proposed a method called SUIM-Net models that implement a fully convolutional encoder-decoder architecture with skip connections between mirrored composite layers. $SUIM - Net_{VGG}$ is the SUIM-Net model variant that utilizes 12 encoding layers of a pre-trained VGG-16 whereas $SUIM - Net_{RSB}$ consists of multiple residual skip blocks. Their method was prepared to classify five different object categories based on the SUIM dataset in which $SUIM - Net_{VGG}$ achieved an accuracy of 84.14% and $SUIM - Net_{RSB}$ achieved 77.77%.

Another method for underwater segmentation was done on synthetic underwater images by implementing SegNet architecture and Support Vector Machine (SVM) to classify three different object categories which it achieved an accuracy of 87% [9]. SegNet was used to first perform image segmentation which the authors stated that the objects in the images were not cleanly segmented. Hence, they applied SVM to further classify the pixels to improve the size and shape characteristics of the segmented objects.

Underwater image segmentation method based on DeepLabV3+ was proposed by [10]. It was prepared for 16 different object categories from a homemade dataset [10]. However, DeepLabV3+ had multiple issues such as target classification error or target pixel mixing, unclear boundary segmentation of the target, the contour for segmented

objects were incomplete and the information for the features available in an image was insufficient [10]. To reduce such issues, the implementation of unsupervised colour correction (UCM) was added in the encoding structure of the method [10]. UCM was used to improve the overall quality of the original image to improve colour correction and reduce lighting problems underwater image. Based on that, the proposed method achieved an accuracy of 64%.

In [11], the author proposed to use Pyramid Scene Parsing Network (PSPNet) to perform semantic segmentation based on multiple domains of datasets. The datasets consist of Cityscapes for urban driving scenes, SUIM for underwater scenes and SUN RGB-D for indoor scenes [11]. PSPNet was chosen due to the model being available with good hardware performance to allow more experimentation [11]. As a result, the author stated that the accuracy of the proposed method on an individual domain can be further improved by training on multiple domains. As such, for the SUIM dataset alone, the proposed method achieved an accuracy of 60.79% to detect all eight object categories available from the dataset.

2.2 Weighted Average Ensemble

Weighted Average Ensemble is a method to reduce the total errors of multiple models by aggregating the predictions from said models [12]. In other words, the predictions from multiple models will be combined based on different weights. For example, for a model the performed the best at classifying Class A but performed the worst at Class B and C will have a lower weight when compared with another model that would perform better at Class B and C and have a lower performance for Class A. When these two models are combined using a weighted average ensemble, a more normalised predicted output is obtained, which, in theory, improves accuracy while lowering total errors in the final prediction when compared to the individual models' output.

3 Methodology

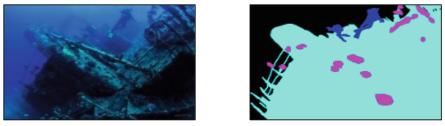
In this section, a method was proposed to perform underwater image segmentation through the combination of multiple deep learning models with different encoders based on weighted average ensemble.

3.1 Dataset

To achieve underwater image semantic segmentation, the dataset that was chosen for this paper was the SUIM dataset [8]. The dataset consists of underwater images which are separated as shown in Table 1, which shows that the SUIM dataset consists of images of different formats and usages. Additionally, the images do not follow a fixed size. The dataset itself had also been already separated for the training dataset by and the test dataset. Figure 1 shows the RGB image and its corresponding ground truth image. The ground truth images are colour coded for each of the segmentation labels which are shown in Table 2.

Table 1. Separation of the SUIM dataset.

	Type	Number	Format
Training data	RGB	1525	.JPG
	Ground truth	1525	.BMP
Testing data	RGB	110	.JPG
	Ground truth	110	.BMP



(a) (b)

Fig. 1. Example of images from the SUIM dataset: **a** the original image; **b** the ground-truth.

Table 2. RGB colour code for each object category.

Object Category	**RGB**	**Code**	**Label**
Background (waterbody)	000	BW	0
Human divers	001	HD	1
Aquatic plants and seagrass	010	PF	2
Wrecks and ruins	011	WR	3
Robots (AUVs/ROVs/instruments)	100	RO	4
Reefs and invertebrates	101	RI	5
Fish and vertebrates	110	FV	6
Seafloor and rocks	111	SR	7

3.2 Image Enhancement

As the nature of underwater images are mostly low lit and have a green-ish or blue-ish hue, the images can be enhanced to improve the overall quality of the images. One of the methods to perform image enhancement was through UCM as shown in [10]. In this project, the UCM algorithm was provided by [13] which implemented traditional image processing techniques such as RGB equalization, stretching and others. Hence,

Table 3. Model names abbreviation.

Encoder	Model	Abbreviation
Inception-V3	FPN	IF
	LinkNet	IL
	U-Net	IU
ResNet34	FPN	RF
	LinkNet	RL
	U-Net	RU

the images in the SUIM dataset, were enhanced with UCM and were saved in.PNG format for further usages.

3.3 Proposed Method

The proposed method is shown in Fig. 2. As mentioned before, it consisted of multiple deep learning models (i.e., U-Net, LinkNet and FPN) combined with multiple encoders (i.e., Inception-V3 and ResNet34) using weighted average ensemble. The individual models were then compared with the proposed method to provide comparison and benchmark on the ensemble approach to the single model approach. Table 3 shows the abbreviation of the models used as shown in Fig. 2.

Additionally, the proposed method was not compared with the method developed by the authors of the dataset ($SUIM - Net_{VGG}$ and $SUIM - Net_{RSB}$) [8] because the number of object categories used were different. The proposed method used all eight object categories provided from the dataset whereas the authors in [8] only used five of the object categories. Instead, the proposed method in this paper will be compared with the method proposed by the author in [11] where the author had used the same eight object categories.

3.4 Dataset Preparation

During the data preparation phase of the project, an error was detected in the dataset used, specifically only for the training data. The dataset was detected to have RGB images that did not have matching sizes with its corresponding ground truth images. Hence, these images were removed, and this reduced the size of the training data from 1525 images to 1488 images. However, this did not occur for the test data, which meant there were no further changes done on it.

As the dataset was a small sized dataset, further data augmentation was done to virtually increase its size. For the context of this paper, horizontal flip was the only data augmentation done to the dataset. Shifting was not applied due to possible distortion of image at the borders, which might influence the segmentation results. Hence, the images were only flipped from left to right. Vertical flip was not applied as it would make some of the images to not be realistic such as an upside-down image of a sunken ship. Horizontal

flip was applied on all the images from the training dataset, which doubled its sized from 1488 to 2976 for both the RGB images and ground truth images.

After that, every single image was then resized to an image resolution of 160×160 pixels for the usages of models for U-Net, LinkNet and FPN. The image resolution was chosen due to requirement of the respective models which were provided by the Segmentation Models library [14].

Ground truth images were further processed by resizing them based on the same image resolution mentioned before. After that, grayscale images were created based on the corresponding RGB values as shown in Table 2 from the resized images. Binary masks for each object category were then obtained from the grayscale images to be used for training and testing purposes.

Lastly, the training data was then split into training dataset and validation dataset with training dataset with a ratio of 80:20. There were no further data split for testing purposes as the dataset itself includes a testing set to be used as mentioned before.

3.5 Encoders

The two encoders (i.e., Inception-V3 and ResNet34) were provided by the Segmentation Models library [14]. The RGB images that were used for both training and testing stages were then pre-processed based on the respective encoders.

3.6 Models

The Segmentation Models library was used to obtain each of the models. The models were used with its default parameters, where one of the more prominent parameters was the weight applied on each of the models. The weight used was *ImageNet.* Other than that, the models' optimizer used was the *Adam* optimizer with the learning rate of 0.0001 and *softmax* as its activation function. Other important parameters to train the model were batch size set to 16, epochs to 25, verbose to 1 and steps per epoch to 149. After each of the models were trained, the models were then saved.

3.7 Weighted Average Ensemble

The saved models were then loaded and used to perform segmentation. The segmentation results from the six models of IF, IL, IU, RF, RL, and RU were then combined based on weighted average ensemble. Hyperparameter tuning was employed to decide the best possible weights combination based on accuracy of the result. It was done through exhaustive search where the result of every single combination of weight values that would total to a value of 1.0 were applied on the six models. The combination that achieved the highest Mean Intersection over Union (MIoU) [11] was then recorded and applied on the six previously mentioned models where the output was considered as the final output of the proposed method. Table 4 shows the weights for the models.

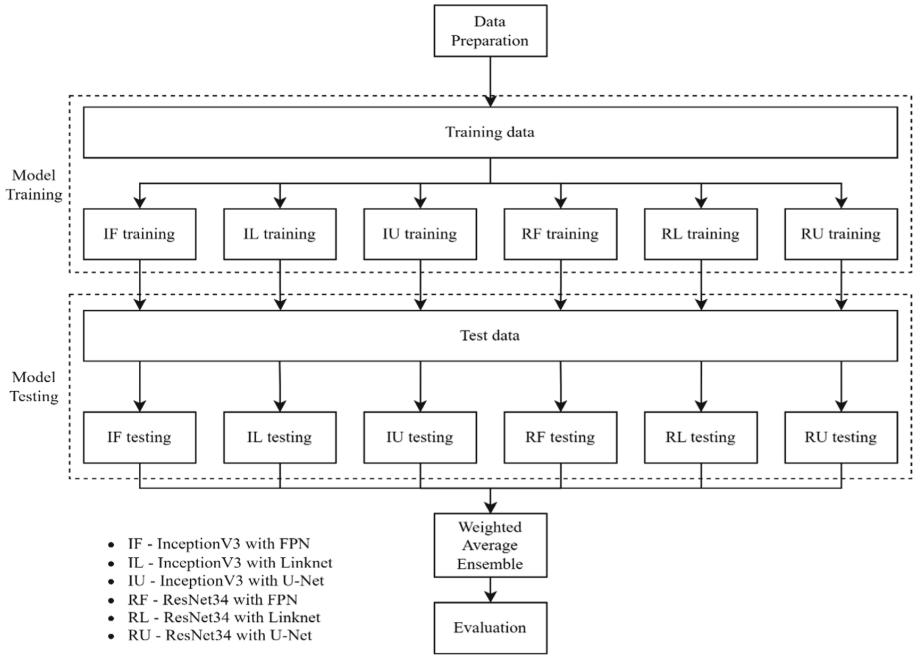


Fig. 2. Structure of the proposed method

Table 4. Recorded weights of the six models.

Models	IF	IL	IU	RF	RL	RU
Weights	0.2	0.1	0.2	0.2	0.1	0.2

4 Results

As previously mentioned, the MIoU metric, as shown in Eq. (1), was used to evaluate the models. MIoU is currently one of the most popular evaluation metrics for semantic segmentation [11]. It calculates the average of Intersection over Union (IoU) of each class present in an image. Hence IoU of each class is equally important and would directly affect the MIoU.

$$MIoU = \frac{1}{N_{class}} \sum_{i=1}^{N_{class}} \frac{TP(i)}{TP(i) + FP(i) + FN(i)} \tag{1}$$

4.1 Quantitative Analysis

Table 5 shows the comparison of the proposed method with the six individual models respectively with the MIoU of the models which also included the result of model

Table 5. Comparison of results.

Model	MIoU (%)
IL	55.49
RL	56.43
RU	56.84
RF	57.84
IU	59.14
IF	59.59
[11]	60.79
Proposed Method	62.48

proposed by [11]. As shown in Table 5, there were significant differences in MIoU of the proposed method to all the individual models, where the proposed method obtained a higher MIoU of 62.48%. Based on the results, the application of weighted average ensemble would perform better when compared to the evaluation of individual models.

4.2 Qualitative Analysis

Figure 3 shows the RGB images, corresponding ground truth images and the segmentation results of each individual model (IL, RL, RU, RF, IU and IF), and the proposed method. For images (a) and (b), the predicted results based on the proposed method looked the most similar to the ground truth image as all of the object categories were able to be detected correctly. This cannot be said when compared with the predicted results of the individual models. This was because the respective predicted results detected extra object categories that were not in the ground truth image. Based on the results, the proposed method was quite robust for images with clear and very distinct differences in textures between the objects in an image.

However, for images (c) and (d), the proposed method's outputs were only able to have some object categories predicted correctly. This can be seen for image (c) where the pixels which were predicted to be RI (Pink) was SR (White) as well as WR (Sky Blue) was not detected at all. This was because of the similarities in terms of textures or colours that were presented in image (c) had reef-like features. Image (d) also faced the same issue where the pixels that were in the class of SR (White) were detected as BW (Black) due to how the area that were represented as SR (White) or sand and sea-floor looked very similar to BW (Black) or background water.

5 Discussion

The result of the proposed method can be considered as above average when referring to its MIoU. It was also higher when compared with one of the methods mentioned before which used the same dataset and number of object categories which was proposed by

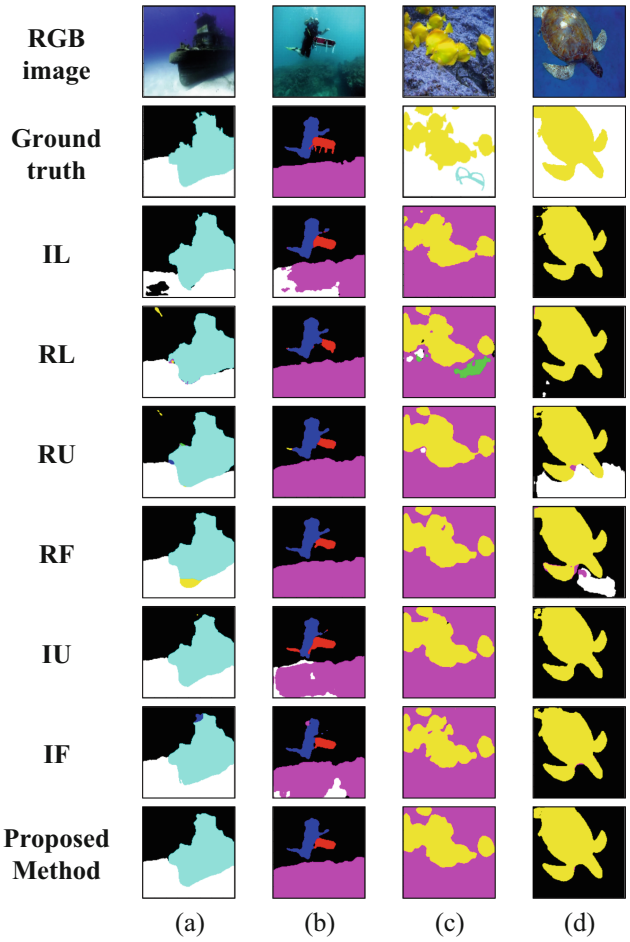


Fig. 3. Underwater image segmentation result on RGB images.

[11]. However, due to some challenges and issues, performance of the proposed method was hampered and could be improved. One of them was the smaller sized dataset that was used to train the model in the proposed method. The training step employed 1488 images, which was a small number even if doubled by applying horizontal flip which increased its size to 2976 images. Considering that deep learning models often require tens or hundreds of thousands of images to be adequately constructed, more images were required to improve on the performance of the proposed method.

Another limitation that impacted the performance of the proposed method was the low resolution of the images. Lower image resolution would mean that an image would lose the precise details of the objects in that image as well as the edges of those objects would be smoother. This would make less obvious details of an object to basically disappear or be modified until it loses its original features. Hence, by using a higher image resolution it would be possible to improve on the performance of the proposed method as the images would be of a higher quality.

Lastly, this paper has shown that underwater image segmentation can be implemented by combining the predicted outputs from multiple models through weighted average ensemble. This was not implemented by the methods proposed by [8–11] where the authors proposed individual models respectively to perform underwater image segmentation. The result of the proposed method by this paper was deemed to be as good or better with the method shown in [11]. This is because both methods were based on the same dataset and the same number of object categories that each method was prepared to detect. Additionally, to improve on the usage of weighted average ensemble, the models that were used can be fine-tuned by performing hyperparameter tuning to find the best parameters settings of each individual model.

6 Conclusion

Underwater image segmentation using deep learning techniques can be a very important tool in the marine industry, specifically underwater exploration. Given the different deep learning techniques that have been implemented and compared, a strategy that combines the results of many models would be deemed effective. This can be observed in the paper's findings, which show that combining the predicted outputs of several models using weighted average ensemble would achieve a higher accuracy when compared to the predicted output from individual models. Further work can be done by using larger sized datasets, experiment on higher image resolution and fine-tuning the models.

Acknowledgments. All authors would like to show great appreciation towards Multimedia University for the support provided to conduct this project.

Authors' Contributions. Study conception and design, analysis and interpretation of results: Muhammad Hidayat Jauhari, Noramiza Hashim;
Draft manuscript preparation: Muhammad Hidayat Jauhari;
All authors reviewed the results and approved the final version of the manuscript.

References

1. *Ocean Exploration and Its Importance*. (n.d.). https://oceanexplorer.noaa.gov/backmatter/whatisexploration.html
2. Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–22. DOI: https://doi.org/https://doi.org/10.1109/TPAMI.2021.3059968
3. Ronneberger, O., Fischer, P., & Brox, T. (2017). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 1–8. DOI: https://doi.org/https://doi.org/10.1007/978-3-319-24574-4_28
4. Chaurasia, A., & Culurciello, E. (2017). *LinkNet : Exploiting Encoder Representations for Efficient Semantic Segmentation*. DOI: https://doi.org/https://doi.org/10.1109/VCIP.2017.8305148

5. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). *Feature Pyramid Networks for Object Detection.* DOI: https://doi.org/10.48550/arXiv.1612.03144
6. He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition.* DOI: https://doi.org/10.48550/arXiv.1512.03385
7. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 2818–2826. DOI: https://doi.org/10.1109/CVPR.2016.308
8. Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., Enan, S. S., & Sattar, J. (2020). Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. *IEEE International Conference on Intelligent Robots and Systems*, 1769–1776. DOI: https://doi.org/https://doi.org/10.1109/IROS45743.2020.9340821
9. O'Byrne, M., Pakrashi, V., Schoefs, F., & Ghosh, B. (2018). *Semantic Segmentation of Underwater Imagery Using Deep Networks Trained on Synthetic Imagery.* 1–15. DOI: https://doi.org/10.3390/jmse6030093
10. Liu, F., & Fang, M. (2020). *Semantic Segmentation of Underwater Images Based on Improved Deeplab.* DOI: https://doi.org/https://doi.org/10.3390/jmse8030188
11. Naber, F. (2021). *Semantic Segmentation on Multiple Visual Domains. June.* DOI: https://doi.org/10.48550/arXiv.2107.04326
12. Jiang, J. (2020, October 14). *Simple Weighted Average Ensemble | Machine Learning | by Jinhang Jiang | Analytics Vidhya | Medium.* https://medium.com/analytics-vidhya/simple-weighted-average-ensemble-machine-learning-777824852426
13. Wang, Y., Song, W., Fortino, G., Qi, L. Z., Zhang, W., & Liotta, A. (2019). An Experimental-Based Review of Image Enhancement and Image Restoration Methods for Underwater Imaging. *IEEE Access*, *7*, 140233–140251. DOI: https://doi.org/https://doi.org/10.1109/ACCESS.2019.2932130
14. Yakubovskiy, P. (2019). *Segmentation Models.* GitHub. https://github.com/qubvel/segmentation_models