FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 4 | Pages 1209-1218 | e-ISSN: 2347-470X

Deep Learning for Enhanced Marine Vision: Object Detection in Underwater Environments

Radhwan Adnan Dakhil[1*] and Ali Retha Hasoon Khayeat[2]

[1]*Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq*
[2]*Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq, ali.r@uokerbala.edu.iq*

***Correspondence:** Radhwan Adnan; radhwan.a@s.uokerbala.edu.iq

ABSTRACT- This study leverages the Semantic Segmentation of Underwater Imagery (SUIM) dataset, encompassing over 1,500 meticulously annotated images that delineate eight distinct object categories. These categories encompass a diverse array, ranging from vertebrate fish and invertebrate reefs to aquatic vegetation, wreckage, human divers, robots, and the seafloor. The use of this dataset involves a methodical synthesis of data through extensive oceanic expeditions and collaborative experiments, featuring both human participants and robots. The research extends its scope to evaluating cutting-edge semantic segmentation techniques, employing established metrics to gauge their performance comprehensively. Additionally, we introduce a fully convolutional encoder-decoder model designed with a dual purpose: delivering competitive performance and computational efficiency. Notably, this model boasts a remarkable accuracy of 88%, underscoring its proficiency in underwater image segmentation. Furthermore, this model's integration within the autonomy pipeline of visually-guided underwater robots presents its tangible applicability. Its rapid end-to-end inference capability addresses the exigencies of real-time decision-making, vital for autonomous systems. This study elucidates the model's practical benefits across diverse applications like visual serving, saliency prediction, and intricate scene comprehension. Crucially, the utilization of the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) elevates image quality, enriching the foundation upon which our model's success rests. This research establishes a solid groundwork for future exploration in underwater robot vision by presenting the model and the benchmark dataset.

Keywords: Underwater Object Detection; Deep Learning; Convolutional Neural Network (CNN); Underwater Imaging; Image Enhancement.

ARTICLE INFORMATION

Author(s): Radhwan Adnan Dakhil and Ali Retha Hasoon Khayeat;
Received: 0/0/0; **Accepted**: 02/12/2023; **Published**: 26/12/2023;
e-ISSN: 2347-470X;
Paper Id: IJEER EMS 588;
Citation: 10.37391/IJEER.110443
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-11/ijeer-110443.html

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Object detection and image segmentation are essential techniques in studying marine life, enabling researchers to gain insights into underwater ecosystems[1]. However, underwater images often suffer from degradation due to light attenuation in water, making the extraction of meaningful information through segmentation a challenging task.

In recent years, Underwater Object Detection (UOD) has emerged as a prominent area in computer vision and image processing. UOD focuses on identifying visually distinctive and semantically meaningful objects in underwater images, separating them from the background. This allows for a better understanding of marine organisms and their interactions within their environment.
Saliency detection, a key component of UOD, has been extensively studied across various disciplines, including computer vision, neuroscience, robotics, and graphics. It involves identifying the most visually striking regions in an image by analyzing features such as contrast, color, spatial information, and texture[2]. This enables the detection of salient objects and helps researchers isolate them from the background. However, the segmentation of underwater images presents unique challenges due to the degradation of image quality caused by light attenuation [3]. Overcoming these challenges is crucial for obtaining accurate and meaningful segmentation results, leading to a better understanding of marine life and the underwater environment.

Semantic segmentation and scene parsing for visually-guided underwater robots lag behind existing solutions in other domains. This is primarily due to two practical limitations. Firstly, underwater imagery exhibits distinct visual characteristics, including domain-specific object categories, unique background patterns, and optical distortion artifacts. Consequently [4], state-of-the-art models trained on terrestrial data are not directly applicable to underwater scenes. Secondly, the absence of comprehensive underwater datasets hinders large-scale training and benchmarking of semantic segmentation models for general-purpose use. Existing datasets are often specific to particular applications, such as coral-reef classification [5, 6] or fish detection [7, 8], and lack the diversity and breadth required for broader research. Moreover, traditional class-agnostic approaches are limited to simpler tasks like foreground segmentation or obstacle detection, and they do not generalize well to multi-object semantic segmentation.

To address the aforementioned limitations, we have chosen to utilize the SUIM dataset, which serves as a large-scale annotated resource specifically designed for semantic segmentation in general-purpose robotic applications [9-11]. This dataset offers an extensive collection of object categories, such as fish, reefs, aquatic plants, and wrecks/ruins, that hold significant interest in the context of underwater exploration and surveying. Additionally, it provides essential pixel-level annotations for human divers, robots/instruments, and sea-floor/rocks, which are crucial for facilitating human-robot cooperative applications [12, 13]. The SUIM dataset comprises a total of 1525 natural underwater images, each meticulously paired with corresponding ground truth semantic labels, ensuring the availability of accurate training data for our model. Furthermore, to comprehensively evaluate the generalization capacity of our approach, the dataset includes a distinct test set consisting of 110 images that have not been used during the model training process.

By leveraging the richness and comprehensiveness of the SUIM dataset, our study aims to significantly enhance the accuracy and robustness of the semantic segmentation model for underwater imagery, making noteworthy contributions to advancements in the field of underwater robotics and exploration.

In this research paper, we delve into the field of Underwater Object Detection, exploring novel methodologies and techniques for accurately identifying and extracting objects from underwater images. We aim to contribute to the advancement of marine research by developing robust algorithms that address the specific challenges posed by underwater imagery. By leveraging recent advancements in computer vision and image processing, we strive to improve the accuracy and efficiency of UOD systems, enabling researchers to study marine organisms with greater precision. Through our research, we hope to deepen our understanding of marine ecosystems and contribute to the conservation and management of our underwater world.

2. RELATED WORKS

Underwater detection tasks have a longstanding history of employing machine learning algorithms. Traditional methodologies have leaned on handcrafted features for underwater object detection, encompassing shape, color, and texture characteristics. For instance, in a study referenced as [10], texture and color features were amalgamated with Support Vector Machines (SVM) to discern underwater corals across various scales. Convolutional Neural Networks (CNNs) were employed by Choi [14] for the classification of fish species, addressing real-time detection requirements. Yang et al. [15] achieved real-time underwater object detection using the YOLOv3 framework [14]. Li et al. [16]utilized the Fast-RCNN framework for fish species detection, subsequently enhancing the speed of fish detection with Faster-RCNN [17]. Villon et al. [18] utilized a deep learning model for the detection of coral reef fishes. Real-time detection demands were also fulfilled by Yang et al. [19], employing the YOLOv3 framework [20] for underwater object detection. Additionally, [18]employed the Fast-RCNN framework for fish species detection, later adopting Faster-RCNN [21] to optimize fish detection speed. Chuang et al. [12] employed texture features extracted using phase Fourier transform for fish detection, while other algorithms incorporated more sophisticated features such as Scale-Invariant Feature Transform (SIFT) [13]and Histogram of Oriented Gradients (HOG) [22].

The utilization of handcrafted features, however, had inherent limitations. Firstly, their task-specific nature impeded generalization, as features tailored for scenes with weak illumination might be ill-suited for well-illuminated underwater environments or scenarios involving substantial changes in the objects to be detected. Secondly, the disjointed nature of feature extraction and classification often led to suboptimal performance, as exemplified by Villon et al. [22], who demonstrated lagging performance in fish classification using HOG features with SVM, falling behind end-to-end deep learning frameworks. Furthermore, proposing and validating effective handcrafted features demanded significant domain expertise.

In contrast, supervised deep learning algorithms have demonstrated the capacity to autonomously derive features from extensive datasets. Deep learning, as a specialized subset of machine learning, employs layered structures inspired by biological neural networks for data analysis. This approach necessitates substantial training data from which it extracts useful and discriminative features with minimal human intervention [23]. Unlike traditional machine learning models, which are often task-specific and require human adjustments, deep learning architectures effectively learn features directly from input data. Deep learning networks have showcased remarkable performance in various computer vision tasks, including image classification, segmentation, object detection, and tracking, and have been widely deployed in underwater object detection.

Despite the advantages of deep learning-based detection models over traditional machine learning models, challenges persist. Deep learning models may encounter difficulties with noisy data and class imbalance, leading to challenges in effectively detecting small objects and resulting in high false positives and false negatives. Consequently, ongoing efforts are essential to address these challenges in deep learning-based underwater object detection. Additionally, Kim et al. [11]proposed a method based on multi-template object selection and color-based image segmentation within the broader context of underwater object detection.

3. SEMANTIC SEGMENTATION

Semantic segmentation for underwater object detection is a challenging computer vision task that involves the accurate classification and delineation of various objects and regions within underwater imagery [16]. It plays a critical role in understanding the complex underwater environment and has significant applications in marine research, environmental monitoring, underwater robotics, and ocean exploration [17, 19].

In the context of underwater object detection, the goal of semantic segmentation is to partition an input underwater image into distinct semantic regions, where each pixel is assigned a specific object category label. Unlike object detection, which focuses on recognizing and localizing individual objects within an image, semantic segmentation provides a more fine-grained understanding of the scene by assigning meaningful labels to every pixel [10], thereby facilitating a pixel-wise analysis of the underwater environment [20]. To achieve semantic segmentation for underwater object detection, deep learning-based approaches have emerged as state-of-the-art techniques. Convolutional Neural Networks (CNNs) serve as the foundation for these methodologies due to their ability to automatically learn hierarchical features from images. Fully Convolutional Networks (FCNs) are a popular choice for this task, as they are designed specifically for dense pixel-wise predictions and allow end-to-end learning.

The process of semantic segmentation begins with the acquisition of a sufficiently large and diverse dataset of underwater images, each manually annotated with pixel-level ground-truth labels corresponding to the different object categories present, such as corals, fish, rocks, sand, and other marine organisms or structures. During training, the deep learning model is fed with the annotated data to learn to identify relevant features that characterize each object category. The model is optimized to minimize the pixel-wise classification loss, ensuring accurate predictions for each pixel's semantic label.

In the inference phase, the trained model is applied to new, unseen underwater images. The model processes the input image and outputs a pixel-wise probability map, where each pixel is associated with the likelihood of belonging to a specific object category [24]. A thresholding step is often applied to obtain the final segmentation mask, where each pixel is assigned the label of the most probable object category.

However, the complex nature of underwater imagery poses several challenges for semantic segmentation. Underwater images are prone to degradation due to absorption, scattering, and color attenuation, leading to reduced visibility and image quality. Moreover, the presence of unique underwater artifacts, such as backscatter and noise, can hinder accurate object detection.

4. THE SUIM DATASET

The SUIM dataset encompasses a range of object categories that are crucial for semantic labeling in underwater imagery analysis. These categories include waterbody background (BW), human divers (HD), aquatic plants/flora (PF), wrecks/ruins (WR), robots and instruments (RO), reefs and other invertebrates (RI), fish and other vertebrates (FV), and sea-floor and rocks (SR) [25]. To represent these categories within the image space, a 3-bit binary RGB color coding scheme is employed, as illustrated in *table 1*.

Table 1: Object Categories and Corresponding Color Codes in the SUIM Dataset

Object category	RGB color Code	RGB color Code
Background (waterbody)	000	BW
Human divers	001	HD
Aquatic plants and sea-grass	010	PF
Wrecks or ruins	011	WR
Robots (AUVs/ROVs/instruments)	100	RO
Reefs and invertebrates	101	RI
Fish and vertebrates	110	FV
Sea-floor and rocks	111	SR

For training and validation purposes, the SUIM dataset comprises a total of 1525 RGB images, while an additional set of 110 test images is provided to facilitate benchmark evaluation of semantic segmentation models. These images exhibit diverse spatial resolutions, including dimensions such as 1906 × 1080, 1280 × 720, 640 × 480, and 256 × 256. Careful selection of these images was conducted, drawing from a large collection gathered during oceanic explorations and human-robot collaborative experiments conducted in various water environments. In addition, a small subset of images from existing large-scale datasets, namely EUVP [4], USR-248 [26], and UFO-120 [27], was incorporated to introduce a diverse range of natural underwater scenes and experimental setups for human-robot cooperation. The population distribution of each object category, their pairwise correlations, and the distributions of RGB channel intensity values within the SUIM dataset are visualized in *figure 1*.

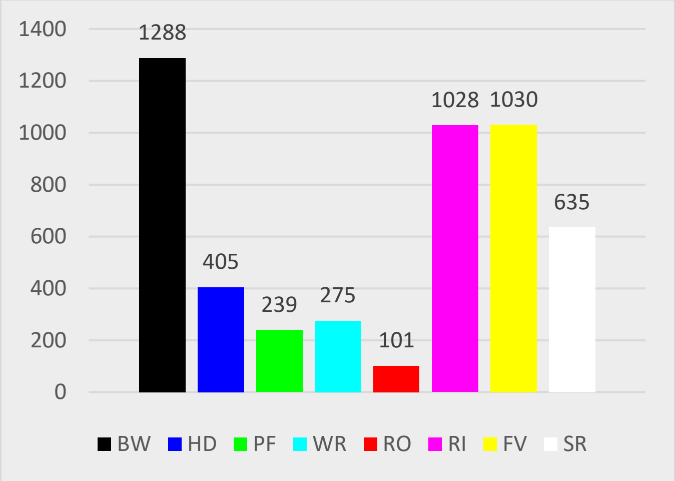


Figure 1: Statistics of Object Categories Values in the SUIM Dataset

The pixel annotations of all images in the SUIM dataset were performed by seven human participants. Figure 2 provides a glimpse of some sample images with their corresponding pixel annotations. To ensure consistent classification of potentially confusing objects, such as plants/reefs and vertebrates/invertebrates, we adhered to the guidelines outlined in [28] and [29]. These guidelines helped ensure accurate and reliable labeling of objects of interest within the dataset.

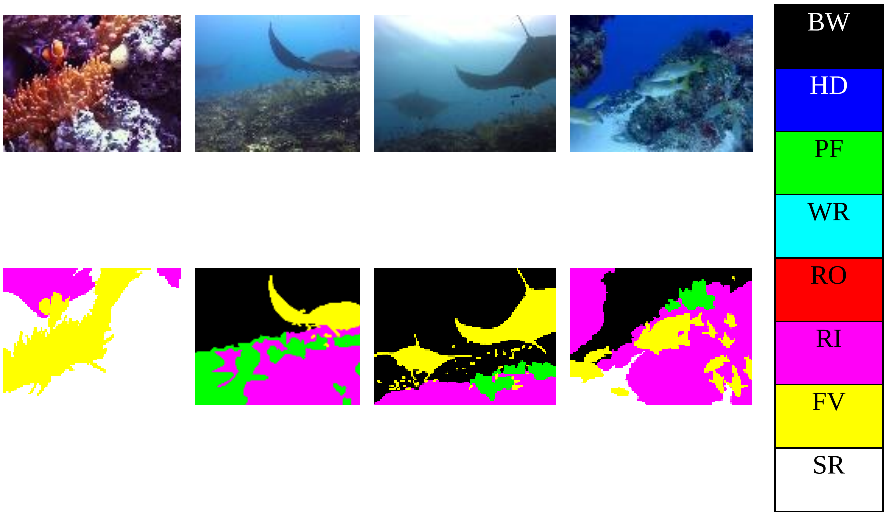


Figure 2: Sample Images and Corresponding Pixel Annotations in the SUIM Dataset

5. PRE-PROCESSING

Images captured under non-uniform light conditions may suffer from color attenuation, scattering effects, and low contrast, leading to a loss of information content. Schettini and Corchs [30] provided an overview of previous research on underwater image enhancement to address this issue and preserve lost information. Among the various degradation aspects, contrast loss significantly impacts classification performance. To ensure consistent image quality and enhance contrast, we have incorporated various pre-processing sub-steps as follows**:**

5.1 Image Super Resolution using ESRGAN

Image super-resolution is the process of enhancing the resolution and quality of a low-resolution image [29]. In underwater imaging, this preprocessing step is crucial due to challenges that often lead to low-quality and low-resolution images [30]. ESRGAN (Enhanced Super-Resolution Generative Adversarial Networks) is a state-of-the-art deep learning method for image super-resolution. It employs a GAN (Generative Adversarial Network) architecture, with a generator network creating high-resolution images and a discriminator network distinguishing between generated and ground truth high-resolution images [31].

To utilize ESRGAN for underwater image super-resolution, a large-scale dataset of high-quality underwater images is collected and used for model training. The ESRGAN model learns the mapping from low-resolution to high-resolution underwater images, taking into account unique underwater image features, such as light scattering and absorption-induced blur, to generate visually pleasing and informative high-resolution images [32]. Once trained, the ESRGAN model can be applied to enhance the resolution of new underwater images, greatly benefiting various underwater applications, including object detection and classification. Refer to *figure 3* for the ESRGAN architecture.

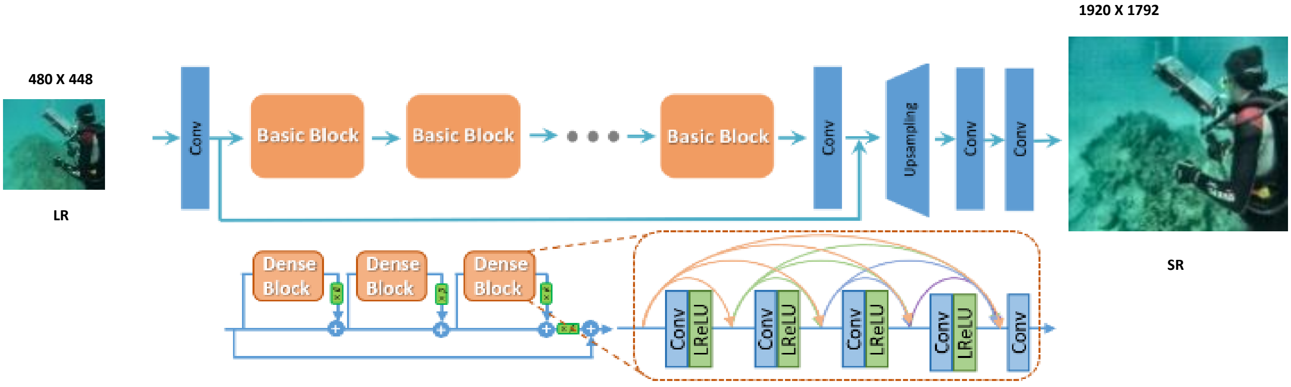


Figure 3: Architecture of Enhanced Super-Resolution Generative Adversarial Networks (ERSGAN)[31]

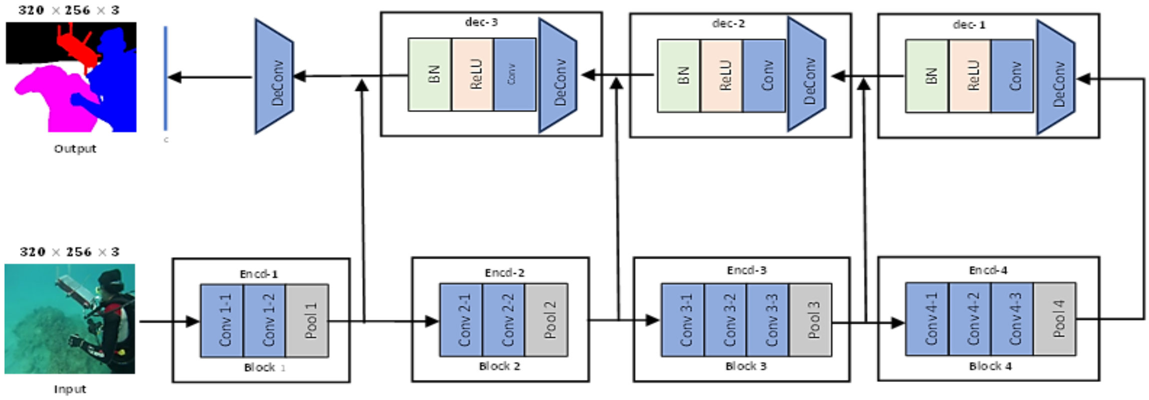


Figure 4: Architecture of the Proposed End-to-End Model for Semantic Segmentation in Underwater Images

The model utilizes the initial four blocks of a pre-trained VGG-16 model for encoding, and subsequently employs three mirrored decoder blocks along with a deconvolution layer for decoding and generating the semantic segmentation map.

6. PROPOSED METHODOLOGY

Underwater image processing is a challenging task due to environmental factors that often degrade image quality. This below pseudocode introduces a systematic approach to address these challenges, leveraging techniques such as super-resolution, data augmentation, and morphological operations for robust object detection in underwater scenarios.

```
    Input: images
    Output: mask,evaluation_metrics
    # read image from training folder
    for i= to NoTrainingImage(= 1525)
     image=dataset_training_folder(i)
     # Super resolution each image using ESRGAN
     Super_ resolution_image=ESRGAN(image)
     # Apply augmentation for each image.
     augmented_data =augmentation(Super_
    resolution_image,
     'rotation_range': 0.2,'width_shift_range':
     0.05,'height_shift_range': 0.05,  'shear_range': 0.05,
     'zoom_range': 0.05, 'horizontal_flip': True, 'fill_mode':
     'nearest')
    # resize each image for training using our model
     resized_images(i) = resize (augmented_data, (320,
 240, 3))
    end
    # Model Initialization
    model = our_model(base, resized_images(all),
    n_classes=5)
    # Training and Testing Loop
    epoch = 1
    best_loss = 0
    while epoch <= 50:
     if training_mode:
      loss = train_model(model, train_data_generator)
      if epoch > 1 and loss < best_loss:
       # Save Trained Model
        save_model(model)
          best_loss = loss
       # Load Trained Model for Testing
        trained_model = load_trained_model()
       # Create Folders for all Classes (5 folders)
        create_class_folders()
       # Test all images in the testing folder
       for j=1 to NoTestingImage(=110)
        image=dataset_testing_folder(j)
        Super_ resolution_image=ESRGAN(image)
        resized_images(j) = resize (Super_ resolution_image,
 (320, 240, 3))
       Predicated_mask=test_model(trained_model,image(j))
        # Morphological Operation
        mask=morphological_operations(Predicated_mask)
        # Performance Evaluation
        evaluation_metrics = evaluate_performance(mask)
        save(mask)
 end
```

6.1 Network Architecture

Our primary focus lies in elevating the performance of our model, which leverages a neural network with 12 encoding layers obtained from pre-training. A visual representation of the architecture details can be found in *figure 4*. The central goal of our research centers around attaining enhanced outcomes and results through this model.

The strategy we have delineated is illustrated in *figure 5* and has been formulated based on a thorough exploration of pertinent literature as well as an exhaustive study of existing techniques and models. This comprehensive literature review encompassed a comparative analysis of diverse models concerning image contrast enhancement, image segmentation, and salient object detection. The proposed methodology encompasses the subsequent stages:

6.1.1. Initial Preprocessing: Underwater Image Super-Resolution

The first step involves enhancing the resolution of underwater images. Several super-resolution models were considered, and after thorough evaluation, ESRGAN [31] was selected as the most suitable approach for our super-resolution process.

6.1.2. Semantic Segmentation Model: Fully Convolutional Encoder-Decoder Architecture

Our proposed model employs a fully convolutional encoder-decoder architecture with skip connections between mirrored composite layers. This model is designed to perform semantic segmentation on the enhanced images, facilitating object detection and classification.

6.1.3. Post-Processing: Morphological Operations

After semantic segmentation, we apply morphological operations as a post-processing step. These operations help refine the segmentation results, making them more accurate and enhancing the quality of the images.

6.1.4. Performance Evaluation: F-score and Intersection over Union (IOU)

To assess the effectiveness of our methodology, we evaluate the results using two metrics: F-score and IOU. These metrics provide insights into the accuracy and quality of object detection and segmentation in the processed images.

It's important to underscore that the theoretical justification for the effectiveness of our proposed technique is rooted in the architectural design choices, preprocessing stages, and the specific model components we employ. This effectiveness can be attributed to the well-considered integration of these elements, allowing our model to excel in terms of performance and outcomes, which aligns with the central objective of our research.

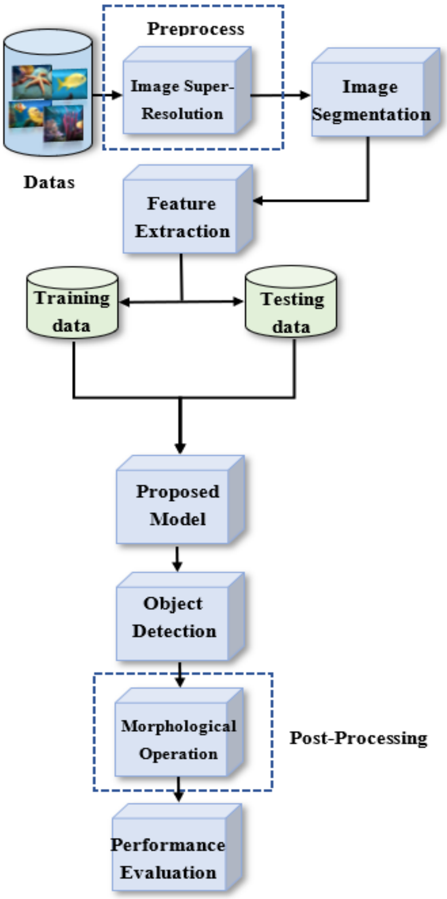


Figure 5: Block Diagram for Underwater Image

6.2 Training Pipeline and Implementation Details

The present study addresses the task of establishing a mapping from the input domain X, which comprises natural underwater images, to their corresponding semantic labeling Y in RGB space. To achieve this mapping, an end-to-end training approach is adopted, where the neural network is trained to minimize the cross-entropy loss [33] by comparing the predicted pixel labels with the ground truth pixel labels. This training strategy is aimed at enabling the network to perform semantic segmentation effectively, resulting in accurate and meaningful pixel-wise predictions in the RGB space for the underwater images. The training settings for semantic segmentation are outlined in *table 2* These settings encompass the hardware utilized, training resolution, parameters to mitigate overfitting, learning rate, image augmentation techniques, saving of training parameters, and maximum iterations.

The training pipeline is implemented using TensorFlow libraries [34] on a Windows host equipped with a Nvidia GTX 1080 graphics card. For optimization, we utilize the Adam optimizer [35] with a learning rate of 10^{-4} and a momentum of 0.5. These settings enable iterative learning to improve the network's performance over time. To enhance the training process and improve generalization, we apply various image transformations as part of data augmentation during training. These transformations help introduce diversity and variability in the training data, contributing to better model robustness and performance.

By formulating the problem as a supervised learning task and utilizing the aforementioned training pipeline and implementation details, we aim to train a model that effectively maps natural underwater images to their corresponding pixel-level semantic labels in RGB space.

Table 2: Underwater Object Detection Training Settings

Category	Configuration item	Configuration value
1. Network	Deep learning network	CNN
2. Hardware	GPU card used	Nvidia GTX 1080
3. Training Resolution	Image resolution during training	1906 × 1080, 1280 × 720, 640 × 480, and 256 × 256 pixels
4. Learning Rate Adjustment	Learning rate	0.0001
5. Image Augmentations	rotation range	0.2
	Width shift range	0.05
	Height shift range	0.05
	Zoom range	0.05
	Horizontal flip	enabled
6. Epoch Number	Number of Epochs	50
7. Data Saving	Save data every	5,000 Iteration
8. Total Training Iterations	Total Iterations	250,000 Iteration

7. RESULTS AND DISCUSSION

We discussed the SUIM dataset and its various use cases for semantic segmentation and saliency prediction in *section 3*. To evaluate the performance of state-of-the-art (SOTA) models, adopted two distinct training configurations, which are described in detail below:

7.1. Semantic Segmentation with Five Major Object Categories

The dataset comprises five major object categories, namely HD, WR, RO, RI, and FV. All other objects in the dataset are considered as background and are represented by the color (000) RGB. To perform semantic segmentation, each model was designed to produce five channels of output, with one channel dedicated to each of the major object categories. These separate pixel masks were then combined to create RGB masks, facilitating visualization of the segmentation results. The primary objective of this configuration was to enable the models to accurately classify and segment input images into the specified five object categories (see *figure 6*).

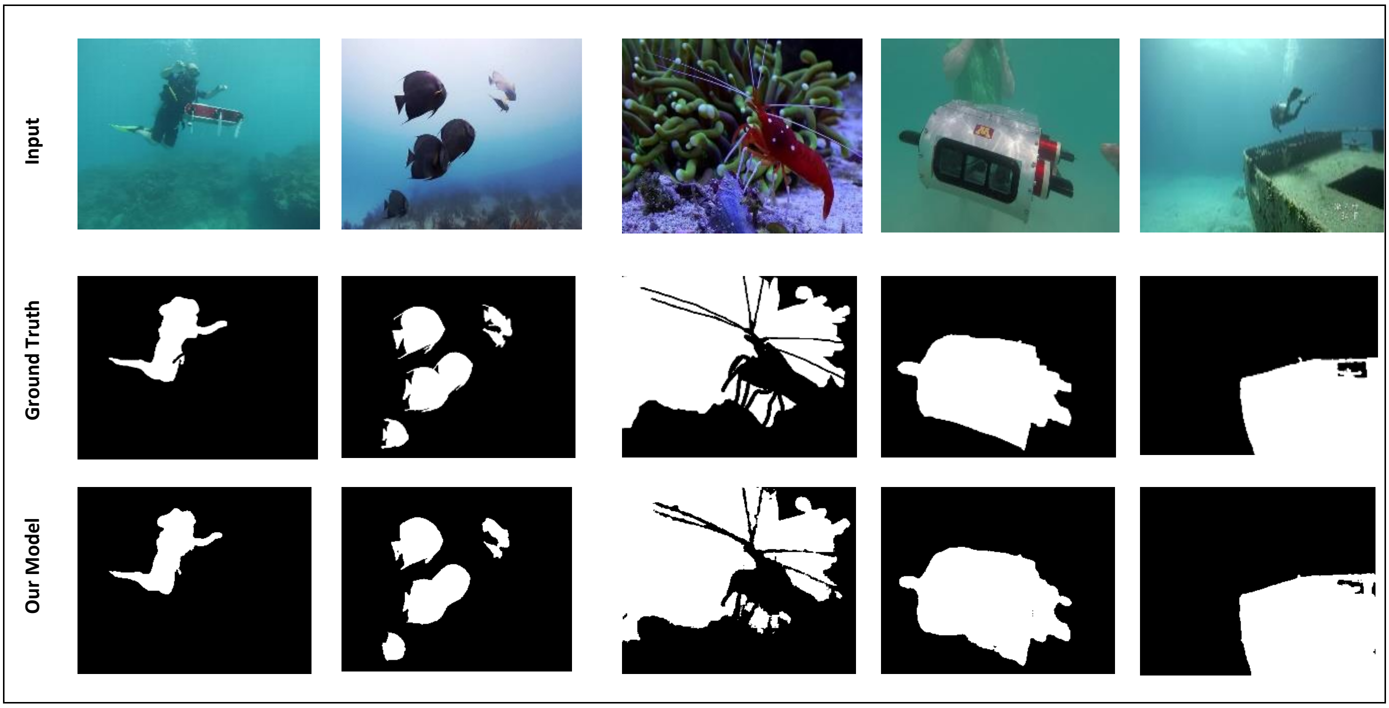


Figure 6: Qualitative Comparison of Semantic Segmentation with Object Categories (HD, WR, RO, RI, and FV) Using Our Model and Ground Truth

7.2. Single-Channel Saliency Prediction

In this specific setup, the focus was on predicting saliency regions within the input images. To achieve this, the ground truth intensities of pixels belonging to the HD, RO, FV, and WR categories were set to 1.0, while pixels corresponding to all other categories were set to 0.0. During training, the models were tasked with predicting a single-channel output representing saliency values. Subsequently, the output was thresholded, yielding binary images that depicted the salient regions. This configuration aimed to assess the models' ability to accurately predict areas of interest within the images.

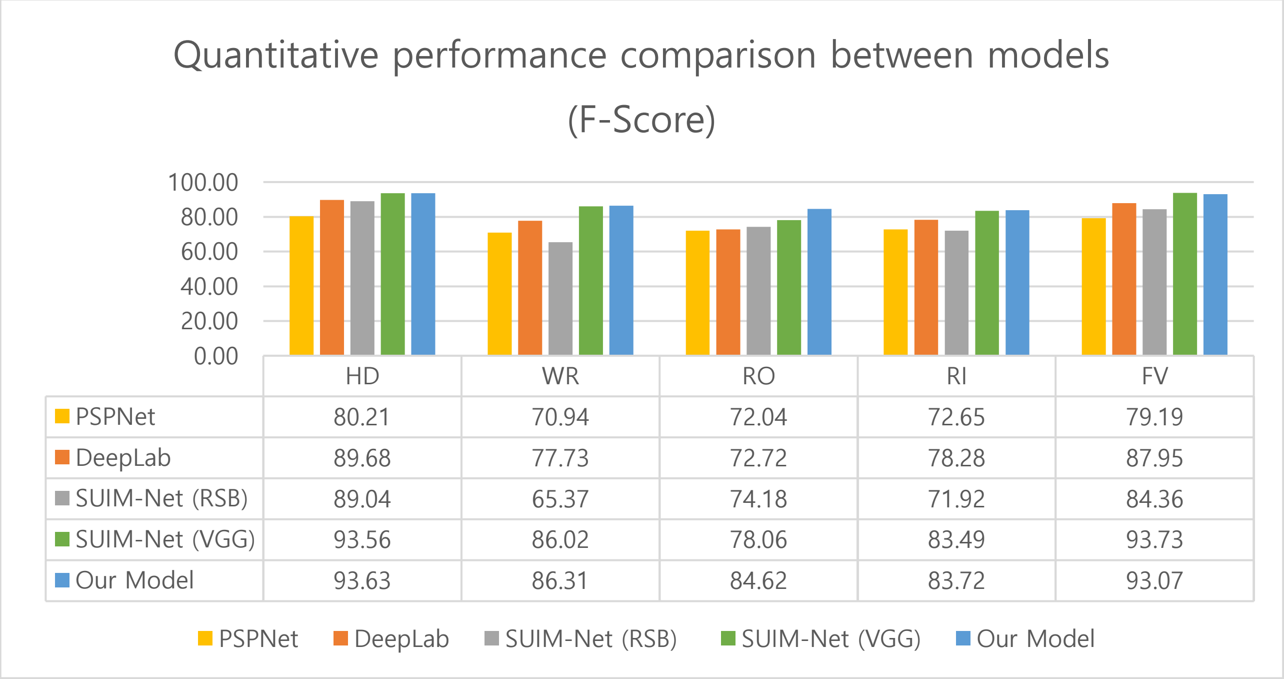
In our evaluation, we compared the performance of all models using standard metrics for region similarity and contour accuracy. The region similarity was measured using the F score (dice coefficient), considering both precision and recall.

$$F = \frac{(2\times P\times R)}{(P+R)} \qquad (1)$$

For contour accuracy, we used the mean IOU (intersection over union) scores, assessing the extent of overlap between predicted and ground truth masks. These well-established metrics allowed us to objectively evaluate the models' segmentation and boundary localization capabilities on the SUIM dataset for semantic segmentation and saliency prediction tasks.

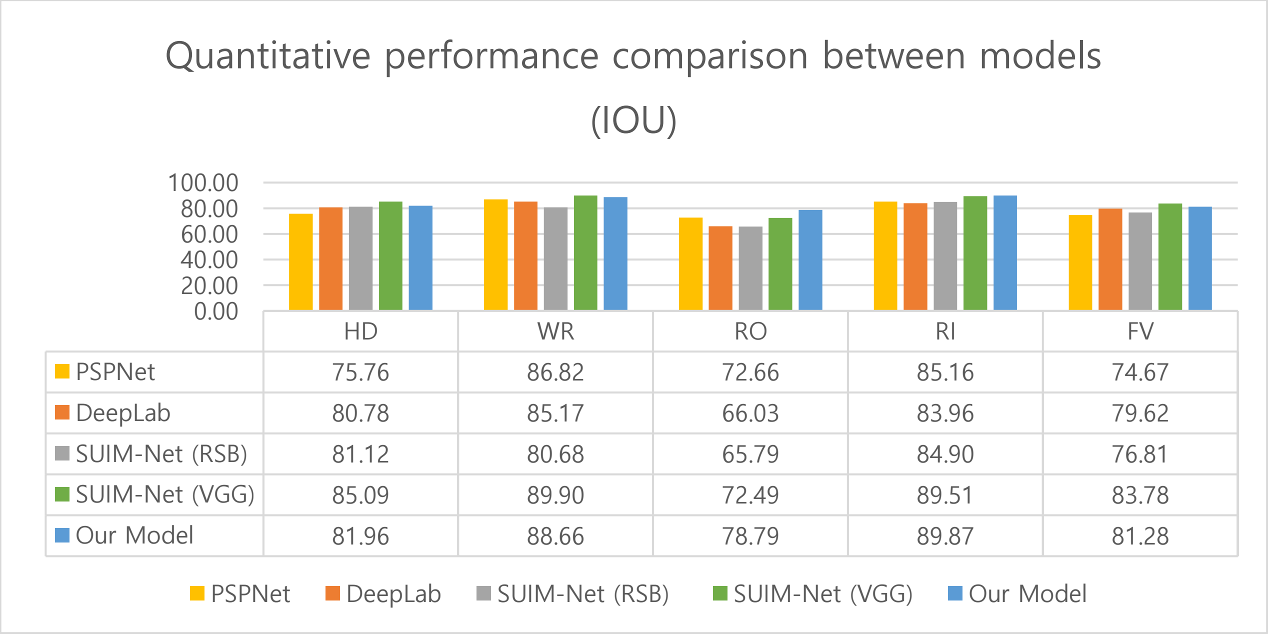
$$\text{IOU} = \frac{(\text{Area of overlap})}{(\text{Area of union})} \qquad (2)$$

The quantitative results presented in *figure 7* and *figure 8* reveal a comparative analysis of F-Score and mIOU scores for semantic segmentation across each object class, as well as saliency prediction scores. Among the evaluated models, DeepLabV3 consistently outperforms others, exhibiting the three highest F-Score and mIOU scores for both semantic segmentation and saliency prediction tasks. Notably, PSPNetMobileNet also delivers competitive results; however, its performance appears to vary across different object classes. In contrast, SUIM-NetRSB and SUIM-NetVGG models, demonstrates consistent and competitive performance in terms of region similarity and object localization.



	HD	WR	RO	RI	FV
🟨 PSPNet	80.21	70.94	72.04	72.65	79.19
🟧 DeepLab	89.68	77.73	72.72	78.28	87.95
⬜ SUIM-Net (RSB)	89.04	65.37	74.18	71.92	84.36
🟩 SUIM-Net (VGG)	93.56	86.02	78.06	83.49	93.73
🟦 Our Model	93.63	86.31	84.62	83.72	93.07

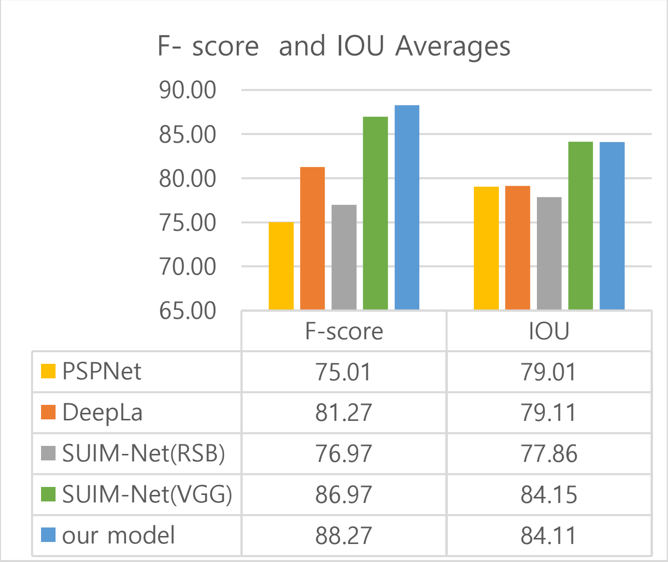
Figure 7: Quantitative performance comparison between models to show the F-Score.



	HD	WR	RO	RI	FV
🟨 PSPNet	75.76	86.82	72.66	85.16	74.67
🟧 DeepLab	80.78	85.17	66.03	83.96	79.62
⬜ SUIM-Net (RSB)	81.12	80.68	65.79	84.90	76.81
🟩 SUIM-Net (VGG)	85.09	89.90	72.49	89.51	83.78
🟦 Our Model	81.96	88.66	78.79	89.87	81.28

Figure 8: Quantitative performance comparison between models to show the IOU

Figure 9 provides a comprehensive comparative analysis, detailing the average F-score and Intersection over Union (IOU) between our model and established techniques in underwater semantic segmentation. Notably, our model's F-score of 88.27 signifies a substantial performance leap over PSPNet (75.01), DeepLa (81.27), SUIM-Net (RSB) (76.97), and SUIM-Net (VGG) (86.97). This numerical advantage aligns with the visual representation in the figure, highlighting our model's proficiency in accuracy and precise object boundary localization. Specifically, our model showcases remarkable accuracy improvements for certain objects, demonstrating adaptability and targeted effectiveness. In comparison, PSPNet falls short in fine-grained segmentation, and DeepLa is surpassed in scene interpretation capabilities. While SUIM-Net (RSB) and SUIM-Net (VGG) exhibit competitive performance, our model emerges as the superior choice. Despite these strengths, considerations for potential limitations, such as generalizability and computational complexity, underscore the need for future research directions to refine these aspects and explore the integration of emerging technologies. This nuanced analysis positions our model as an advanced and versatile solution in the realm of underwater semantic segmentation, with far-reaching implications for diverse applications in underwater image analysis.



	F-score	IOU
🟨 PSPNet	75.01	79.01
🟧 DeepLa	81.27	79.11
⬜ SUIM-Net(RSB)	76.97	77.86
🟩 SUIM-Net(VGG)	86.97	84.15
🟦 our model	88.27	84.11

Figure 9: The Average of F-Score and IOU

8. CONCLUSION

This research addressed the crucial requirements of semantic segmentation and pixel-level detection of salient objects in underwater scenes to empower visually-guided Autonomous Underwater Vehicles (AUVs). Despite the rapid advancements in terrestrial domain literature, existing solutions have been limited by either application-specific nature or outdated methodologies. To overcome these limitations, we present the first large-scale annotated dataset, SUIM, specifically designed for general-purpose semantic segmentation of underwater scenes. The dataset comprises 1525 images with pixel annotations for eight object categories, including fish, reefs, plants, wrecks/ruins, humans, robots, sea-floor/sand, and waterbody background. Additionally, we provide a comprehensive benchmark evaluation of state-of-the-art (SOTA) semantic segmentation approaches on the dataset's test set.

Furthermore, we introduce our model, a fully-convolutional encoder-decoder architecture. Our model demonstrates competitive semantic segmentation performance while offering significantly faster runtime compared to existing SOTA approaches. This delicate balance between robust performance and computational efficiency renders our model suitable for near real-time utilization in attention modeling and servoing tasks for visually-guided underwater robots.

The effectiveness of our approach is reinforced by the achieved result of 88% accuracy in semantic segmentation. This result substantiates the superiority of our model over other methodologies, affirming its capability to accurately detect and classify objects in challenging underwater environments. In achieving these results, we employed Image Super Resolution using ESRGAN as a preprocessing step, effectively enhancing the resolution and quality of low-resolution underwater images. Additionally, morphological operations were utilized to further refine the segmentation results.

The availability of the SUIM dataset and the performance of our model open up new possibilities for various underwater applications. In the near future, we plan to extend the utilization of the SUIM dataset to explore different learning-based models, such as visual question answering and guided search. Our objective is to assess their feasibility in underwater human-robot cooperative applications, thus contributing to the advancement of underwater robotics and exploration.

This research represents a significant step towards bridging the gap in semantic segmentation and object detection methodologies between terrestrial and underwater domains. The SUIM dataset and our efficient model pave the way for enhanced capabilities and practical use of visually-guided AUVs in underwater exploration, marine research, and environmental monitoring. The success of our approach demonstrates the potential for further advancements in underwater computer vision, facilitating progress in the understanding and preservation of underwater ecosystems and marine resources.

REFERENCES

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. J. a. p. a. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017. https://doi.org/10.48550/arXiv.1704.06857.

[2] M. Jian, Q. Qi, J. Dong, Y. Yin, K.-M. J. J. o. v. c. Lam, and i. representation, "Integrating QDWD with pattern distinctness and local contrast for underwater saliency detection," vol. 53, pp. 31-41, 2018. https://doi.org/10.1016/j.jvcir.2018.03.008.

[3] M. Sharma, J. Lim, and H. J. A. S. Lee, "The amalgamation of the object detection and semantic segmentation for steel surface defect detection," vol. 12, no. 12, p. 6004, 2022. https://doi.org/10.3390/app12126004.

[4] M. J. Islam, Y. Xia, J. J. I. R. Sattar, and A. Letters, "Fast underwater image enhancement for improved visual perception," vol. 5, no. 2, pp. 3227-3234, 2020. https://doi.org/10.1109/LRA.2020.2974710.

[5] I. Alonso, M. Yuval, G. Eyal, T. Treibitz, and A. C. J. J. o. F. R. Murillo, "CoralSeg: Learning coral segmentation from sparse annotations," vol. 36, no. 8, pp. 1456-1477, 2019. https://doi.org/10.1002/rob.21915.

[6] A. Haider, M. Arsalan, J. Choi, H. Sultan, and K. R. J. F. i. M. S. Park, "Robust segmentation of underwater fish based on multi-level feature accumulation," vol. 9, p. 1010565, 2022. https://doi.org/10.3389/fmars.2022.1010565.

[7] S. Girija, A. Akhila, D. Deepthi, R. U. Kiran, and P. A. Krishna, "Saliency and Transmission Feature Extraction from Underwater Images Using Level Set Method," in 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, pp. 1-7: IEEE. https://doi.org/10.1109/ICEEICT53079.2022.9768472.

[8] Y. Girdhar, P. Giguere, and G. J. T. I. J. o. R. R. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," vol. 33, no. 4, pp. 645-657, 2014. https://doi.org/10.1177/02783649135073.

[9] L. T. Parker IV, N. Gage, G. Van Anne, C. Tomaszewski, W. Newcomb, and A. Spears, "mTITAN: multi-domain tactical intelligent teaming and autonomous navigation," in Open Architecture/Open Business Model Net-Centric Systems and Defense Transformation 2023, 2023, vol. 12544, pp. 55-64: SPIE. https://doi.org/10.1117/12.2663907.

[10] J. Chamberlain, A. Garcia Seco De Herrera, A. Campello, and A. Clark, "ImageCLEFcoral task: coral reef image annotation and localisation," in CEUR Workshop Proceedings, 2022, vol. 3180, pp. 1318-1328: CEUR Workshop Proceedings.

[11] D. Kim, D. Lee, H. Myung, and H.-T. J. I. S. R. Choi, "Artificial landmark-based underwater localization for AUVs using weighted template matching," vol. 7, pp. 175-184, 2014. https://doi.org/10.1007/s11370-014-0153-y.

[12] M.-C. Chuang, J.-N. Hwang, and K. J. I. T. o. I. P. Williams, "A feature learning and object recognition framework for underwater fish images," vol. 25, no. 4, pp. 1862-1872, 2016. https://doi.org/10.48550/arXiv.1603.01696.

[13] S. Y. Alaba et al., "Class-aware fish species recognition using deep learning for an imbalanced dataset," vol. 22, no. 21, p. 8268, 2022. https://doi.org/10.3390/s22218268.

[14] L. Shen, H. Tao, Y. Ni, Y. Wang, V. J. M. S. Stojanovic, and Technology, "Improved YOLOv3 model with feature map cropping for multi-scale road object detection," vol. 34, no. 4, p. 045406, 2023. http://dx.doi.org/10.1088/1361-6501/acb075.

[15] H. Yang, P. Liu, Y. Hu, and J. J. M. T. Fu, "Research on underwater object recognition based on YOLOv3," vol. 27, pp. 1837-1844, 2021. https://doi.org/10.1007/s00542-019-04694-8.

[16] S. Bosse and P. J. E. P. Kasundra, "Robust Underwater Image Classification Using Image Segmentation, CNN, and Dynamic ROI Approximation," vol. 27, no. 1, p. 82, 2022. https://doi.org/10.3390/ecsa-9-13218.

[17] Z. Chen et al., "Underwater sonar image segmentation combining pixel-level and region-level information," vol. 100, p. 107853, 2022. https://doi.org/10.1016/j.compeleceng.2022.107853.

[18] D. Zhao, B. Yang, Y. Dou, and X. Guo, "Underwater fish detection in sonar image based on an improved Faster RCNN," in 2022 9th International Forum on Electrical Engineering and Automation (IFEEA), 2022, pp. 358-363: IEEE. https://doi.org/10.1109/IFEEA57288.2022.10038226.

[19] J. Wang, X. He, F. Shao, G. Lu, R. Hu, and Q. J. P. o. Jiang, "Semantic segmentation method of underwater images based on encoder-decoder architecture," vol. 17, no. 8, p. e0272666, 2022.

[20] Z. Liu et al., "Canet: Context aware network for brain glioma segmentation," vol. 40, no. 7, pp. 1763-1777, 2021.

[21] G. Han, S. Huang, J. Ma, Y. He, and S. J. a. p. a. Chang, "Meta Faster R-CNN: Towards Accurate Few-Shot Object Detection with Attentive Feature Alignment. arXiv 2021." https://doi.org/10.1609/aaai.v36i1.19959.

[22] S. Villon, M. Chaumont, G. Subsol, S. Villéger, T. Claverie, and D. Mouillot, "Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between Deep Learning and HOG+ SVM methods," in International Conference on Advanced Concepts for Intelligent Vision Systems, 2016, pp. 160-171: Springer. https://doi.org/10.1007/978-3-319-48680-2_15.

[23] Y. LeCun, Y. Bengio, and G. J. n. Hinton, "Deep learning," vol. 521, no. 7553, pp. 436-444, 2015. http://dx.doi.org/10.1038/nature14539.

[24] M. Alavianmehr, M. Helfroush, H. Danyali, and A. J. J. o. r.-t. i. p. Tashk, "Butterfly network: a convolutional neural network with a new architecture for multi-scale semantic segmentation of pedestrians," vol. 20, no. 1, p. 9, 2023. https://doi.org/10.1007/s11554-023-01273-z.

[25] R. A. Dakhil and A. R. H. J. a. p. a. Khayeat, "Review On Deep Learning Technique For Underwater Object Detection," 2022. https://doi.org/10.48550/arXiv.2209.10151.

[26] M. J. Islam, S. S. Enan, P. Luo, and J. Sattar, "Underwater image super-resolution using deep residual multipliers," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 900-906: IEEE.

[27] M. J. Islam, P. Luo, and J. J. a. p. a. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," 2020. https://doi.org/10.48550/arXiv.2002.01155.

[28] "Marine Life Encyclopedia " 2001.

[29] X. Wang et al., "Esrgan: Enhanced super-resolution generative adversarial networks," in Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0-0. https://doi.org/10.1007/978-3-030-11021-5_5.

[30] A. J. a. p. a. Aghelan, "Underwater Images Super-Resolution Using Generative Adversarial Network-based Model," 2022. https://doi.org/10.48550/arXiv.2211.03550.

[31] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: Further improving enhanced super-resolution generative adversarial network," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 3637-3641: IEEE. https://doi.org/10.48550/arXiv.2001.08073.

[32] H. Wang et al., "Simultaneous restoration and super-resolution GAN for underwater image enhancement," vol. 10, p. 1162295, 2023. https://doi.org/10.3389/fmars.2023.1162295.

[33] Z. Zhang and M. J. A. i. n. i. p. s. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," vol. 31, 2018. https://doi.org/10.48550/arXiv.1805.07836.

[34] M. Abadi et al., "{TensorFlow}: a system for {Large-Scale} machine learning," in 12th USENIX symposium on operating systems design and implementation (OSDI 16), 2016, pp. 265-283.

[35] D. P. Kingma and J. J. a. p. a. Ba, "Adam: A method for stochastic optimization," 2014. https://doi.org/10.48550/arXiv.1412.6980.