International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Enhancing Marine Vision: Deep Learning-Based Underwater Object Detection

Radhwan Adnan Dakhil [1], Ali Retha Hasoon Khayeat [2]

[1] *Department of Computer Science, Faculty of Computer Science and Information Technology, Kerbala University, Kerbala, Iraq*
[2] *Department of Computer Science, Faculty of Computer Science and Information Technology, Kerbala University, Kerbala, Iraq*

E-mail address: radhwan.a@s.uokerbala.edu.iq, ali.r@uokerbala.edu.iq

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

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 of items, 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: to deliver competitive performance and computational efficiency. Notably, this model boasts a remarkable accuracy of 88%, underscoring its proficiency in underwater image segmentation. This study elucidates the model's practical benefits across diverse applications such as visual serving, saliency prediction, and intricate scene comprehension. Crucially, the utilization of the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) raises 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: Deep Learning, Convolutional Neural Network (CNN), Underwater Object Detection, Underwater Imaging, Image Enhancement

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 distinct 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 face notable challenges that distinguish them from their terrestrial counterparts. These challenges can be attributed to two fundamental practical constraints. Firstly, underwater imagery exhibits unique visual characteristics, including object categories specific to this domain, distinct background patterns, and optical distortion artifacts. Consequently, state-of-the-art models

E-mail: radhwan.a@s.uokerbala.edu.iq, ali.r@uokerbala.edu.iq

trained on terrestrial data cannot be readily applied to the complexity of underwater scenes [4]. Secondly, the absence of comprehensive underwater datasets presents a significant obstacle to large-scale training and benchmarking of general-purpose semantic segmentation models. Current datasets often cater to particular use cases, like the classification of coral reefs [5, 6]. or fish detection [7, 8]. They often lack the diversity and coverage necessary for broader research endeavors. Additionally, conventional class-agnostic approaches are constrained to relatively simpler tasks like foreground segmentation or obstacle detection, falling short when it comes to multi-object semantic segmentation.

To overcome these limitations, our approach leverages the SUIM dataset, purposefully created to serve as a substantial and annotated resource for semantic segmentation in general-purpose underwater robotics applications [9, 11]. This dataset boasts an extensive array of object categories that hold particular significance in the realm of underwater exploration and surveying, including fish, reefs, aquatic plants, and wrecks/ruins. Moreover, it provides crucial pixel-level annotations for various elements within the images, encompassing human divers, robots/instruments, and seafloor/rocks, which essential for supporting human-robot collaboration applications [12, 13]. The SUIM dataset encompasses a total of 1,525 natural underwater images, each meticulously paired with corresponding ground truth semantic labels. This meticulous curation ensures the availability of precise training data for our model. Furthermore, to rigorously evaluate the model's ability to generalize, the dataset incorporates a distinct test set comprising 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. OBJECT DETECTION IN COMPUTER VISION

Object detection is a complex task within computer vision, presenting challenges similar to other assignments in this field. Training for detection and classification occurs concurrently in various image locations. Convolutional models distribute work across these locations, sharing computation loads. However, unlike localization tasks, object detection requires accounting for a background class when no object is present. Both classifying and locating image regions contribute to the challenge of identifying objects. Successful object detection involves understanding how to segment images and determine object locations. Recognizing an object's location aids in understanding its shape while understanding an object's shape assists in pinpointing its location. [14] For instance, features that appear distinct, such as a person's face and attire, might constitute parts of the same object. Nonetheless, comprehending the object's identity remains difficult without first recognizing the object itself.

A. Underwater Object Detection Based on Object Characteristics

Historically, underwater object detection relied on conventional techniques, employing algorithms that emphasized the recognition of contours, shapes, colors, or a combination of these features to identify objects in images and subsequently classify them.

In a specific study [15], an innovative approach was introduced for underwater object detection. This method initiates by determining the presence of an object in an image through color segmentation. This technique is particularly effective in environments where fish imagery is captured within a controlled setting, featuring a known blue background and uniform lighting conditions. In this context, the term "object" primarily refers to fish. The procedure consists of subtracting the blue channel from the red channel, taking advantage of the observation that pixels representing fish objects usually display higher red and lower blue channel values. This operation results in the creation of a mask where object pixels are assigned a value of one, while background pixels receive a value of zero.

Once the presence of an object is confirmed, the approach proceeds to extract its contour using the information derived from the mask. The resulting contour is subsequently simplified through data reduction, often reducing it to a manageable number of points, such as 40. This simplification enables further analysis, where parameters like the normalized length and turn angles between these points are examined to determine if the object corresponds to a fish. If it is indeed identified as a fish, the tracking process is initiated, and the species of the fish is determined. Species identification is accomplished by assessing specific landmark points on the fish and applying a technique known as Turn Angle Distribution

Analysis (TADA). This approach achieved a notable accuracy of 73.3% when tested on a dataset comprising 300 fish images representing six different species.

Although This technique was advanced for its time and yielded promising outcomes due to the controlled environment with optimal lighting conditions, it possesses limitations. It is constrained by its reliance on a specialized setup and its inability to handle scenarios where fish are partially occluded, bent, or subjected to shadows, which often leads to erroneous object detection in real-world settings.

B. Underwater Object Detection Based on Deep Learning

Deep learning has revolutionized object detection, making it more applicable to real-life scenarios. The approach proposed in [14] was limited by its specific setup and lacked real-world suitability. Deep learning has overcome these limitations by autonomously learning from labeled datasets, enabling object identification in diverse positions and enhancing real-life usability.

In [15], a deep learning approach was implemented for fish detection and classification, specifically tailored for underwater imagery. This approach involved a series of steps, commencing with foreground extraction to enhance object detection. Subsequently, the improved images were input into a Convolutional Neural Network (CNN) with two convolutional layers employing distinct kernel sizes. The output from these layers underwent feature pooling and spatial pyramid pooling, facilitating object recognition across different poses. Final classification was achieved through a classifier layer utilizing Support Vector Machines (SVM). The Fish for Knowledge (F4K) dataset, containing 22,370 images representing 23 fish species, was employed. Remarkably, despite using a relatively less complex network, this method achieved an impressive accuracy of 98.57%.

Nevertheless, certain limitations were observed. Foreground extraction faced real-world constraints, particularly due to the presence of non-fish objects. Additionally, the dataset contained images with varying resolutions, standardized to 47 * 47 pixels. Achieving higher resolutions would necessitate a deeper network, impacting processing time. Furthermore, the dataset exhibited an uneven distribution of species, with some having significantly fewer images.

As for [16], another approach was pursued, involving a deep learning object detection algorithm known as Fast R-CNN. The dataset was compiled from the F4K video repository, featuring 12 fish species and offering a more balanced image distribution compared to [15]. Training involved the use of Stochastic Gradient Descent (SGD). The processing times per image for these algorithms were 24.945, 0.311, and 0.273 seconds, respectively, while their mean Average Precision (mAP) values were 81.2%, 81.4%, and 78.9%. It's essential to highlight that this approach had limitations due to the dataset's focus on well-lit and well-posed fish images, which lacked tailored digital image processing for underwater conditions. As a result, the approach's performance may not be as robust when applied to real-world underwater imagery.

3. RELATED WORK

Underwater recognition and detection tasks have a well-established history of employing machine learning algorithms. Traditional methods in this field heavily relied on manually designed features for the detection of underwater objects, which included characteristics like shape, color, and texture. As an example, in [10], the authors harnessed a combination of texture and color features, complemented by Support Vector Machines (SVMs), to identify various scales of underwater corals. Kim et al. [11] introduced a technique centered on multi-template object selection and color-based image segmentation, while Chuang et al. [12] utilized texture features extracted through the phase Fourier transform for fish detection. Some algorithms went even further by incorporating more advanced features, including the Scale-Invariant Feature Transform (SIFT) [13] and the Histogram of Oriented Gradients (HOG) [17]. These techniques were considered the state-of-the-art methods in the field of underwater object detection for a substantial.

However, the applicability of these hand-crafted features had limitations. First, their task-specific nature hindered their capacity for generalization; features tailored for scenes with weak illumination might not suit well-illuminated underwater environments or scenarios involving substantial changes in the objects to be detected. Second, the disjointed nature of feature extraction and classification often led to suboptimal performance, as demonstrated by Villon et al. [17], who used HOG features with SVM for fish classification, lagging behind end-to-end deep learning frameworks. Additionally, proposing and validating effective hand-crafted features would demand significant expertise.

On the other hand, supervised deep learning algorithms have the ability to independently extract features from large datasets. Deep learning, a specialized subset of machine learning, employs layered structures inspired by biological neural networks to analyze data. It requires substantial training data from which it extracts useful and discriminative features with minimal human intervention [18]. Unlike traditional machine learning models that are task-specific and often require human adjustments, deep learning architectures effectively learn features from input data. Deep learning networks have showcased remarkable performance in a wide range of computer vision tasks, encompassing image classification, image segmentation, object detection, and object tracking. In underwater object detection, deep learning has been widely deployed. Choi

[19] used convolutional neural networks (CNNs) for fish species classification, while Villon et al. [17] employed a deep learning model for detecting coral reef fishes. [20] utilized the Fast-RCNN framework for fish species detection, later adopting Faster-RCNN [21] to enhance the speed of fish detection. Real-time detection requirements were met by Yang et al. [21] using the YOLOv3 framework [23] for underwater object detection. Despite the advantages of deep learning-based detection models over traditional machine learning models, challenges persist. Deep learning models may struggle with noisy data and class imbalance, leading to difficulties in effectively detecting small objects, which result in high numbers of false positives and false negatives. Therefore, further efforts are necessary to address these challenging issues in deep learning-based underwater object detection.

4. 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 [25]. 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 [26, 27].

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 [28]. 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 [29]. 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.

5. THE SUIM DATASET

The SUIM dataset is a rich and all-encompassing collection, spanning a spectrum of object categories crucial for the semantic analysis of underwater imagery. These categories, including background waterbody (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 seafloor and rocks (SR) [30], are visually represented using a 3-bit binary RGB color coding scheme, thoughtfully outlined in Table 1.

TABLE I. OBJECT CATEGORIES AND ASSOCIATED 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 the purpose of training and validation, the SUIM dataset comprises a total of 1,525 RGB images. Furthermore, an additional set of 110 test images is generously provided to facilitate the benchmark evaluation of semantic segmentation models. These images span a diverse range of spatial resolutions, including dimensions such as 1906 × 1080, 1280 × 720, 640 × 480, and 256 × 256. The selection of these images was conducted meticulously, drawing from a vast collection gathered

during oceanic explorations and collaborative experiments involving both humans and robots in a myriad of underwater environments.

Moreover, to introduce a wide range of natural underwater scenes and experimental configurations that are suitable for human-robot cooperation, we judiciously incorporated a smaller subset of images sourced from established Extensive datasets, particularly EUVP [4], USR 248 [31], and UFO 120 [32], were drawn upon. These datasets contributed to the variety of object categories, their associations, and the subtleties in RGB channel intensity values within the SUIM dataset. are vividly illustrated in the captivating visual representation featured in Figure 1.

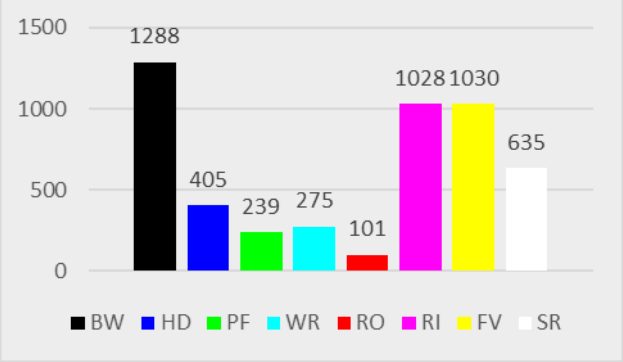


Figure 1. Statistics of Object Categories in SUIM Dataset.

the SUIM dataset stands as a testament to the meticulous work of seven human annotators who dedicated themselves to the intricate task of pixel-level annotations. Figure 2, showcasing these annotations alongside sample images, unequivocally showcases the dataset's exceptional quality and precision.

The paramount objective of this annotation endeavor was to establish consistent object classification throughout the dataset, particularly when faced with potentially confounding distinctions like those between plants/reefs and vertebrates/invertebrates. This stringent approach serves as a guarantee of the dataset's unwavering reliability and its broad applicability in the realms of computer vision and image analysis.

In the pursuit of this precision, we diligently adhered to the guidelines delineated in references [33] and [34]. These invaluable guidelines played a pivotal role in ensuring the accuracy and dependability of object labeling within the dataset, further reinforcing its scholarly and practical value.

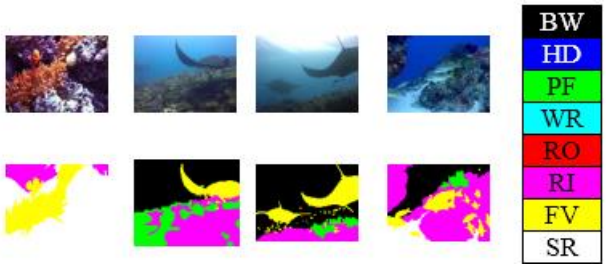


Figure 2. Sample Images and Corresponding Pixel Annotations in the SUIM Dataset.

6. PRE PROCESSING

Images captured in varying or uneven lighting conditions may suffer from color attenuation, scattering effects, and low contrast, leading to a loss of information content. Schettini and Corchs [34] 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 to enhance contrast, we have incorporated various pre-processing sub-steps as follows:

A. Image Super-Resolution using ESRGAN

Image super-resolution is a crucial preprocessing step in underwater imaging, aimed at enhancing the resolution and quality of low-resolution images [35]. Underwater photography often faces challenges that result in low-quality and low-resolution images [36]. Enter ESRGAN, short for Enhanced Super-Resolution Generative Adversarial Networks, a cutting-edge deep learning technique tailored for image super-resolution. It operates on the foundation of a GAN, or Generative Adversarial Network, which comprises a generator network responsible for creating high-resolution images and a discriminator network tasked with distinguishing between generated images and ground truth high-resolution images [37]. Leveraging ESRGAN for underwater image super-resolution begins with the collection of a substantial dataset featuring high-quality underwater images, which serves as the basis for model training. The ESRGAN model, when trained on this dataset, learns the intricate mapping from low-resolution to high-resolution underwater images. Importantly, it takes into account the unique characteristics of underwater images, including challenges like light scattering and absorption-induced blur. By doing so, it produces visually pleasing and informative high-resolution images that are well-suited for underwater applications [38]. Once trained, the ESRGAN model can be applied to elevate the resolution of new underwater images, imparting significant benefits to a variety of underwater applications, particularly those reliant on object detection and

classification. For a visual representation of the ESRGAN architecture, please refer to Figure 3.

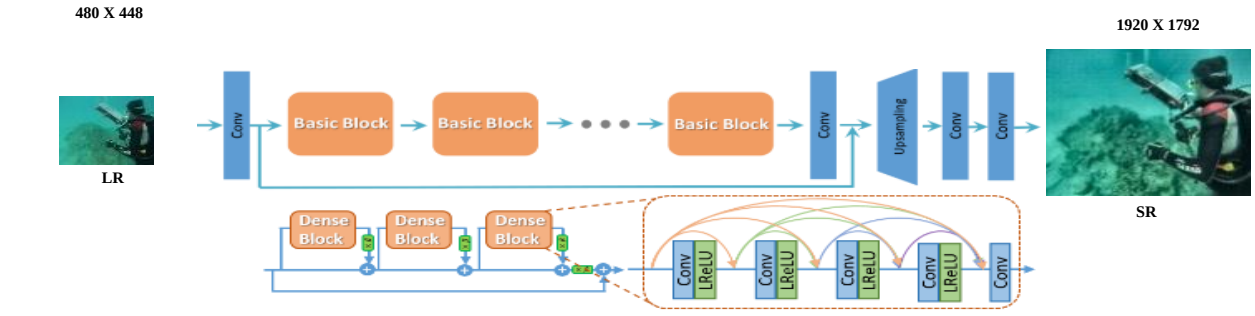


Figure 3. Architecture of Enhanced Super-Resolution Generative Adversarial Networks (ERSGAN) [32].

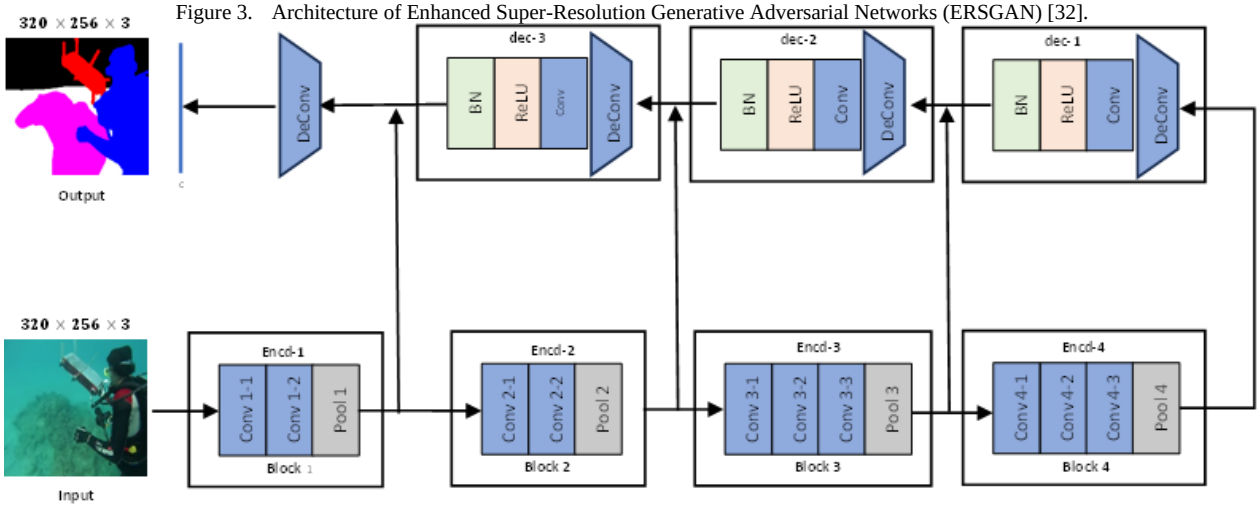


Figure 4. Architecture of the Proposed End-to-End Model for Semantic Segmentation in Underwater Images. The model utilizes the initial four blocks of a pre-trained VGG 16 model for encoding, and subsequently employs three mirrored decoder blocks along with a deconvolution layer for decoding and generating the semantic segmentation map.

7. PROPOSED METHODOLOGY

A. *Network Architecture*

Our primary focus lies in elevating the performance of our model, which leverages a neural network with twelve 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 through this model.

The strategy we have delineated is illustrated in Figure 5 and has been formulated based on a thorough exploration of the 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 comprises a series of carefully considered steps to achieve our research objectives:

1. Initial Preprocessing for Underwater Image Super-Resolution: The first phase of our methodology focuses on enhancing the resolution of underwater images, which often suffer from low quality and resolution. In this regard, we explored several super-resolution models, conducting a thorough evaluation to identify the most suitable approach for our specific needs. Our extensive evaluation led us to select the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) [39] as the optimal solution for our super-resolution process.

2. Model Implementation with Convolutional Encoder-Decoder Architecture: In the next step, we proceeded with the implementation of our model. Our model architecture is based on a fully convolutional encoder-decoder design, featuring skip connections between mirrored composite layers. This architecture is integral to our approach as it plays a crucial role in the extraction and reconstruction of high-resolution information from low-resolution input.

3. **Comparative Assessment of Proposed Solution:** To validate the efficacy of our proposed methodology, we conducted a comparative evaluation against existing models that address similar challenges. This step allows us to quantitatively measure the performance and effectiveness of our approach to other solutions available in the field.

It is essential to highlight that the effectiveness of our proposed technique is grounded in a thoughtful combination of architectural design choices, preprocessing stages, and the specific components of our model. These elements are meticulously integrated to ensure that our model excels in terms of performance and outcomes, aligning with the central objective of our research. By carefully considering these aspects, we aim to contribute a robust and efficient solution for underwater image enhancement and related applications.

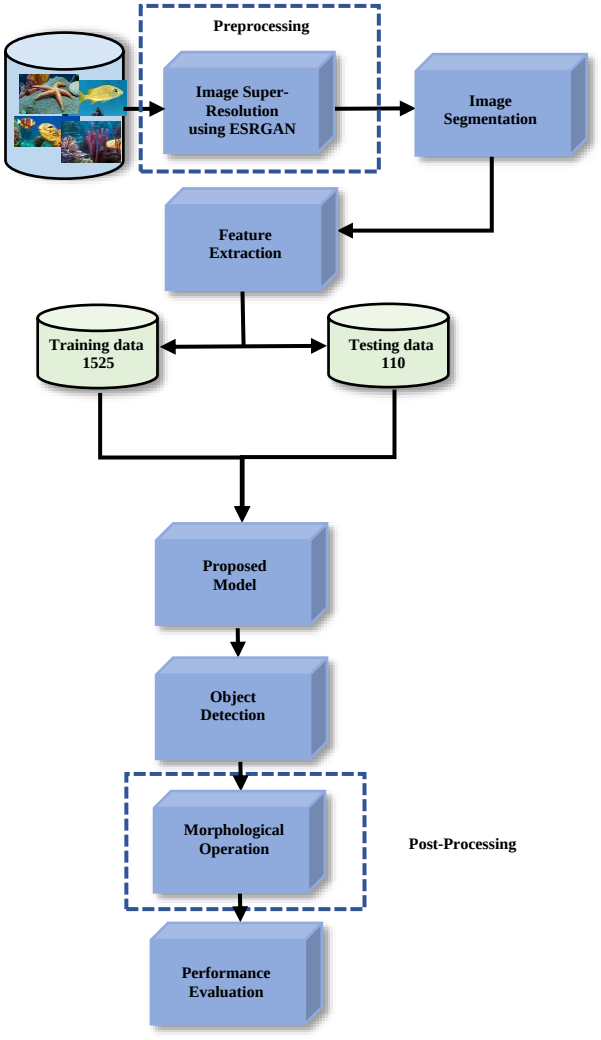


Figure 5. Block Diagram for Underwater Image Object Detection

in terms of performance and outcomes, which aligns with the central objective of our research.

B. Training Pipeline and Implementation Details

In this study, our focus is on establishing a mapping from the input domain X, which consists of natural underwater images, to their corresponding semantic labels Y within the RGB space. To achieve this mapping, we employ an end-to-end training approach, wherein the neural network is trained to minimize the cross-entropy loss [34]. This loss function is utilized to compare the predicted pixel labels with the ground truth pixel labels, enabling the network to effectively perform semantic segmentation. The goal is to generate precise and semantically meaningful pixel-wise predictions in the RGB space for the underwater images.

The training pipeline is implemented using TensorFlow libraries [40] on a Windows host equipped with an NVidia GTX 1080 graphics card. For optimization, we utilize the Adam optimizer [41] 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 to 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 the RGB space. To prevent overfitting, a set of parameters was employed. Simple image augmentation techniques were applied to the dataset, as shown in Table II.

The learning rate, which influences how the optimizer adapts during training, was set to a dynamic value. A large learning rate can lead to rapid changes in weight values, potentially resulting in convergence to a suboptimal solution. Conversely, a low learning rate may cause slow convergence. Therefore, a dynamic learning rate was utilized to ensure a stable gradient and prevent model divergence.

TABLE II. 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	∓ 0.2
	Width_shift	∓ 0.05
	Height_shift	∓ 0.05
	Zoom	∓ 0.05
	Horizontal flip	enabled
6. Data saving	Save data every	5,000 Iteration
7. Maximum training iterations	Maximum iteration	5,000 Iteration

8. RESULTS AND DISCUSSION

To evaluate the performance of state-of-the-art (SOTA) models, we adopted two distinct training configurations, which are described in detail below :

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 to be 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.

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.

During our assessment, we conducted a comparative analysis of all models by employing established metrics to assess the similarity of regions and the accuracy of contours. Specifically, we measured region similarity using the F score (also known as the dice coefficient), which takes into account both precision and recall.

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

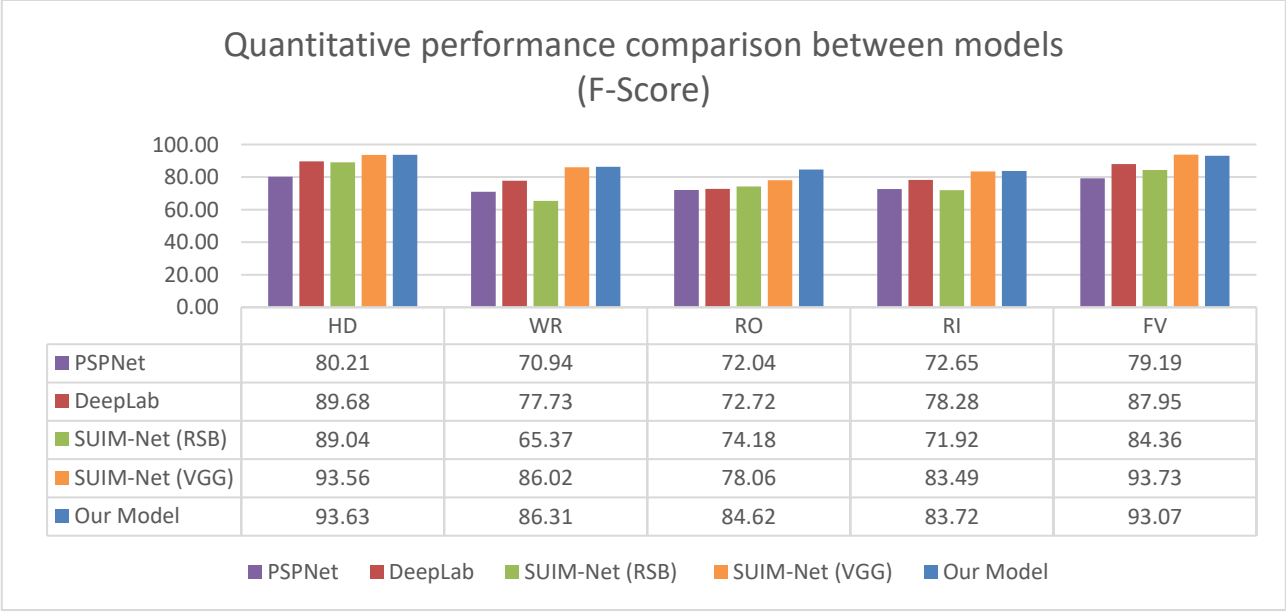
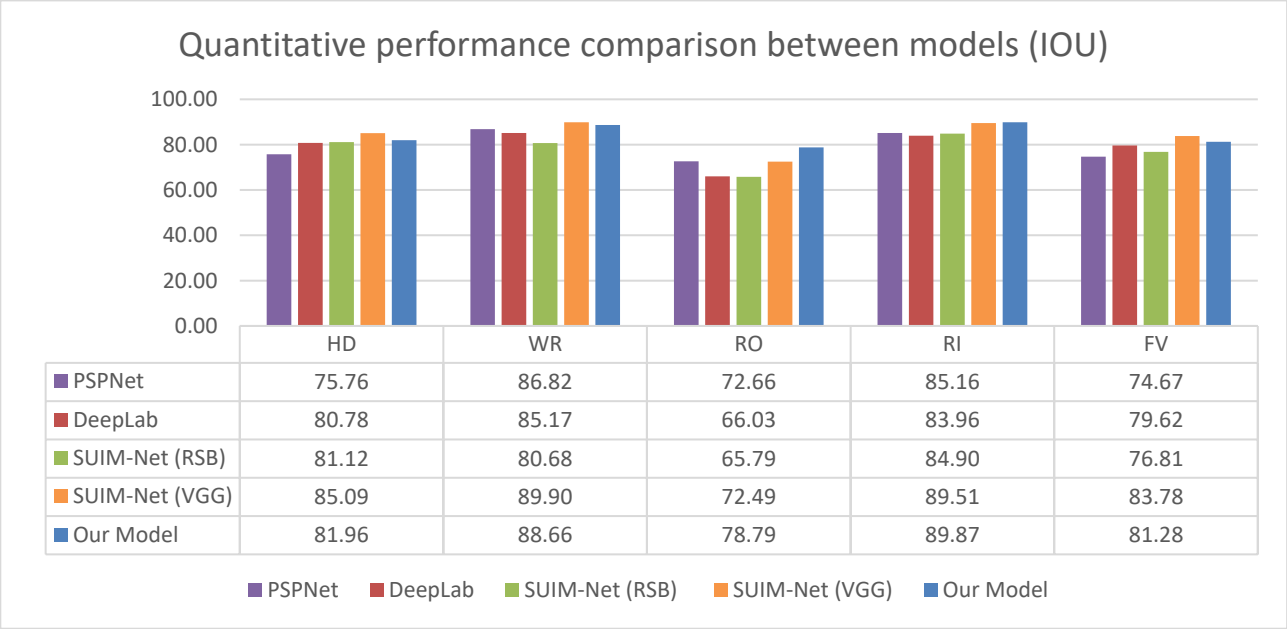


Figure 6. Quantitative performance comparison between models to show the F-Score.



	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

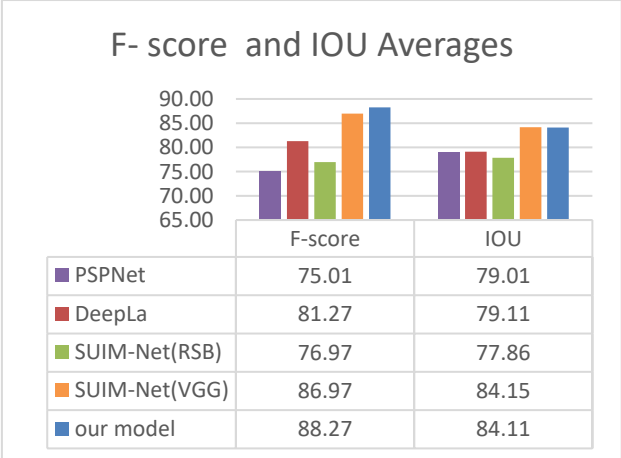
Figure 7. Quantitative performance comparison between models to show the IOU.

Regarding contour accuracy, our evaluation involved utilizing the mean IOU (intersection over union) scores. This measure helped us assess the degree of overlap between the predicted masks and the ground truth masks. These widely recognized metrics were instrumental in providing an objective evaluation of the models' proficiency in semantic segmentation and saliency prediction tasks on the SUIM dataset.

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

The numerical results illustrated in Figures 6 and 7 offer an extensive examination of the F Score and mIOU scores for semantic segmentation across all individual object categories, along with saliency prediction scores. Among the range of models under scrutiny, DeepLabV3 consistently emerges as the top performer, securing the highest three scores for F-Score and IOU in both semantic segmentation and saliency prediction tasks. It's worth noting that PSPNetMobileNet also delivers competitive results, albeit with varying effectiveness across different object categories. In contrast, the SUIM-NetRSB and SUIM-NetVGG models consistently exhibit competitive performance in terms of region similarity and object localization.

Figure 8 shows the average F-score and IOU comparison between our model and the others . Our model outperforms the others in accuracy and object boundary localization, affirming its superiority for underwater semantic segmentation. Interestingly, our model exhibits notable improvements in accuracy for certain specific objects, as illustrated in the accompanying figure . These advancements further underscore the efficacy of our approach and its potential for superior performance in semantic segmentation and saliency prediction tasks compared to the other evaluated models. These advancements further underscore the efficacy of our approach and its potential for superior performance in semantic segmentation and saliency prediction tasks compared to the other evaluated models.



	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 8. Average of F-Score and IOU

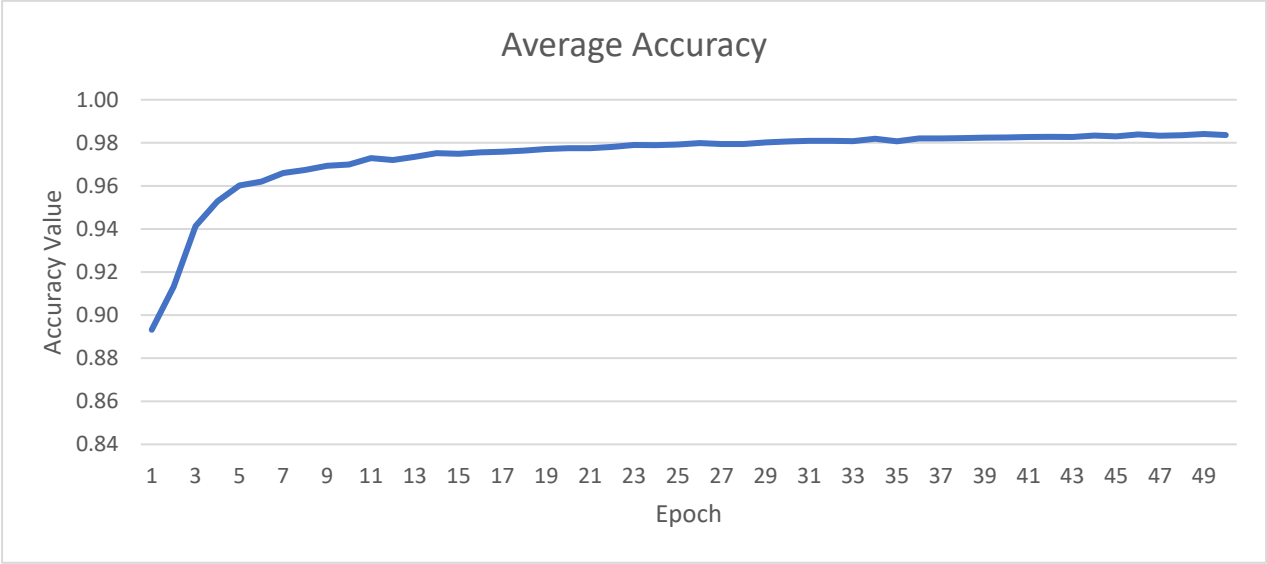


Figure 9. Accuracy versus different epoch plot.

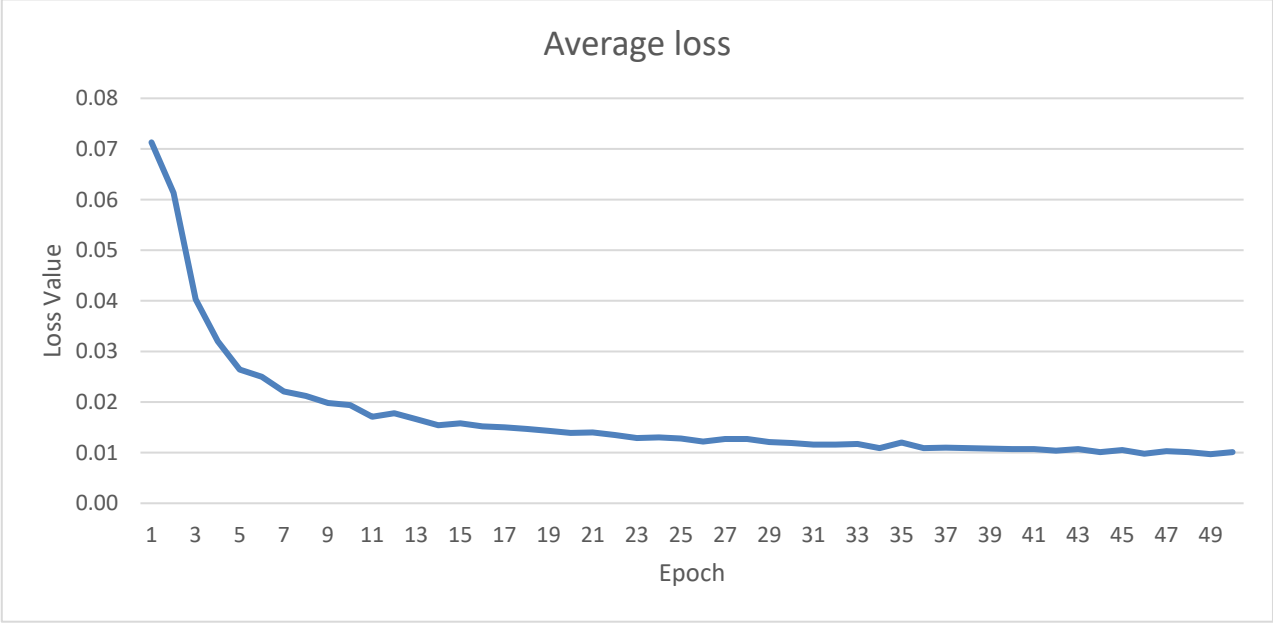


Figure 10. Accuracy versus different epoch plot.

Initiating the training phase, our model leverages a pre-trained model's parameter values as a starting point. Subsequently, the assessment of outcomes commences through the evaluation of the loss function. Typically, this process begins with a notably high loss, gradually diminishing as optimization of the model's parameters unfolds. As evidenced in the training conducted with ERSGAN, the initial average accuracy of 0.8932 ascended commendably to 0.9836, a trend showcased in Figure 9. Simultaneously, the average loss, commencing at 0.0713, exhibited a downward trajectory, culminating at 0.0101, as illustrated in Figure 10. These outcomes affirm the model's adaptability to the underlying challenge.

However, the optimizer's proficiency in problem adaptation and resolution does not inherently guarantee the trained model's precision in object detection. Diverse factors may contribute to this, ranging from an insufficiently diverse training dataset, which may fail to encapsulate all conceivable object scenarios, to potential

limitations rooted in the quality of the image dataset itself. Challenges might arise from inefficiencies in extracting object features due to inadequate image quality, further accentuating the complexity of the detection process.

9. CONCLUSIONS

This study addressed the critical need for semantic segmentation and pixel-level detection of salient objects in underwater environments to enhance the capabilities of visually-guided Autonomous Underwater Vehicles (AUVs). While advancements in terrestrial domains have been well-documented in the literature, existing solutions for underwater scenarios have been constrained by their application-specific nature or outdated methodologies. In response to these limitations, we introduce the SUIM dataset, a pioneering and comprehensive dataset thoughtfully curated for general underwater environment semantic segmentation. This dataset comprises 1,525 images, each meticulously annotated at the pixel level, covering eight distinct object categories, which encompass fish, reefs, plants, wrecks/ruins, humans, robots, seafloor/sand, and waterbody backgrounds. Additionally, we undertake a thorough evaluation of cutting-edge semantic segmentation techniques, employing the dataset's test set for benchmarking purposes.

Our proposed model adopts a fully convolutional encoder-decoder architecture, achieving competitive performance in semantic segmentation while offering significantly improved runtime efficiency compared to existing SOTA approaches. This delicate balance between robust performance and computational efficiency makes our model well-suited for near real-time utilization in tasks such as attention modeling and servoing for visually-guided underwater robots.

Our approach's effectiveness is exemplified by the achievement of an impressive 88% accuracy in semantic segmentation. This remarkable result underscores the superiority of our model when compared to alternative methodologies, clearly demonstrating its ability to accurately detect and classify objects even in challenging underwater conditions. To attain these outstanding results, we thoughtfully incorporated Image Super Resolution using ESRGAN as a preprocessing step, a technique that effectively enhances the resolution and overall quality of low-resolution underwater images. Additionally, we harnessed the power of morphological operations to further refine the segmentation outcomes, ensuring that our model delivers precise and reliable performance.

The release of the SUIM dataset, coupled with the exceptional performance of our model, opens up exciting new opportunities across various underwater applications. In the immediate future, we are eager to harness the full potential of the SUIM dataset to explore diverse learning-based models, such as those geared towards visual question answering and guided searches. Our ultimate objective is to evaluate their feasibility in the context of underwater human-robot collaborative applications. By doing so, we hope to make substantial contributions to the advancement of underwater robotics and exploration, paving the way for cutting-edge developments in these fields.

This research represents a pivotal milestone in narrowing the gap between semantic segmentation and object detection methodologies in terrestrial and underwater domains. With the SUIM dataset at its core and our highly efficient model leading the way, we are opening doors to enhanced capabilities and practical usage of visually-guided Autonomous Underwater Vehicles (AUVs) in the realms of underwater exploration, marine research, and environmental monitoring. The remarkable success of our approach serves as a testament to the potential for further breakthroughs in underwater computer vision, facilitating significant progress in the comprehension and conservation of underwater ecosystems and marine resources.

In the future, we plan to extend and fortify our contributions in the domains of underwater image enhancement and object detection. First and foremost, we are dedicated to constructing a unified enhancement-detection framework, seamlessly amalgamating low-level image enhancement and high-level object detection. By unifying these components, we will eliminate the need for separate models, subsequently mitigating the associated time and resource overhead. This integration holds the promise of delivering superior solutions and significantly expediting the processing time, as it eliminates the necessity for image transmission between distinct models. Additionally, We will primarily concentrate on creating an innovative multi-task loss function designed to thoroughly assess the efficacy of this integrated framework.

Secondly, in response to the growing demand for reduced computational complexity in real-time underwater applications, we will concentrate on the integration of various model compression algorithms into our framework. These techniques, encompassing weight binarization, weight pruning, and compact block design, are geared towards substantially reducing memory and computational overhead, thereby enhancing the efficiency of our framework. Our ultimate objective is to formulate a deep learning compression algorithm that minimizes storage and energy requirements, rendering deep networks suitable for real-time deployment on AUVs and remotely operated vehicles (ROVs). These future research endeavors will further solidify and expand upon our contributions in this field.

REFERENCES

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

[2] Jian M, Qi Q, Dong J, Yin Y, Lam K-MJJovc, representation i. Integrating QDWD with pattern distinctness and local contrast for

underwater saliency detection. 2018;53:31-41. https://doi.org/10.1016/j.jvcir.2018.03.008.

[3] Sharma, M., Lim, J., & Lee, H. (2022). The amalgamation of the object detection and semantic segmentation for steel surface defect detection. Applied Sciences, 12(12), 6004. https://doi.org/10.3390/app12126004.

[4] Islam MJ, Xia Y, Sattar JJIR, Letters A. Fast underwater image enhancement for improved visual perception. 2020;5(2):3227-34. https://doi.org/10.1109/LRA.2020.2974710.

[5] Alonso I, Yuval M, Eyal G, Treibitz T, Murillo ACJJoFR. CoralSeg: Learning coral segmentation from sparse annotations. 2019;36(8):1456-77. https://doi.org/10.1002/rob.21915.

[6] Haider, A., Arsalan, M., Choi, J., Sultan, H., & Park, K. R. (2022). Robust segmentation of underwater fish based on multi-level feature accumulation. Frontiers in Marine Science, 9, 1010565. https://doi.org/10.3389/fmars.2022.1010565.

[7] Girija, S. P., Akhila, A., Deepthi, D., Kiran, R. U., & Krishna, P. A. (2022, February). 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) (pp. 1-7). IEEE. https://doi.org/10.1109/ICEEICT53079.2022.9768472.

[8] Girdhar Y, Giguere P, Dudek GJTIJoRR. Autonomous adaptive exploration using realtime online spatiotemporal topic modeling. 2014;33(4):645-57. https://doi.org/10.1177/02783649135073.

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

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

[11] Kim D, Lee D, Myung H, Choi H-TJISR. Artificial landmark-based underwater localization for AUVs using weighted template matching. 2014;7:175-84. https://doi.org/10.1007/s11370-014-0153-y.

[12] Chuang M-C, Hwang J-N, Williams KJIToIP. A feature learning and object recognition framework for underwater fish images. 2016;25(4):1862-72. https://doi.org/10.48550/arXiv.1603.01696.

[13] Alaba, S. Y., Nabi, M. M., Shah, C., Prior, J., Campbell, M. D., Wallace, F., ... & Moorhead, R. (2022). Class-aware fish species recognition using deep learning for an imbalanced dataset. Sensors, 22(21), 8268. https://doi.org/10.3390/s22218268.

[14] Villon S, Chaumont M, Subsol G, Villéger S, Claverie T, Mouillot D, editors. Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between Deep Learning and HOG+ SVM methods. Advanced Concepts for Intelligent Vision Systems: 17th International Conference, ACIVS 2016, Lecce, Italy, October 24-27, 2016, Proceedings 17; 2016: Springer. https://doi.org/10.1007/978-3-319-48680-2_15.

[15] A. K. Gupta, A. Seal, M. Prasad, and P. J. E. Khanna, "Salient object detection techniques in computer vision—A survey," vol. 22, no. 10, p. 1174, 2020.

[16] [50] N. Chen, W. Liu, R. Bai, and A. J. A. I. R. Chen, "Application of computational intelligence technologies in emergency management: a literature review," vol. 52, pp. 2131-2168, 2019.

[17] [51] H. Qin, X. Li, J. Liang, Y. Peng, and C. J. N. Zhang, "DeepFish: Accurate underwater live fish recognition with a deep architecture," vol. 187, pp. 49-58, 2016.

[18] [52] H. Huang, H. Zhou, X. Yang, L. Zhang, L. Qi, and A.-Y. J. N. Zang, "Faster R-CNN for marine organisms detection and recognition using data augmentation," vol. 337, pp. 372-384, 2019.

[19] LeCun Y, Bengio Y, Hinton GJn. Deep learning. 2015;521(7553):436-44. http://dx.doi.org/10.1038/nature14539.

[20] Aruna, S. K., Deepa, N., & Devi, T. (2023, May). Underwater Fish Identification in Real-Time using Convolutional Neural Network. In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 586-591). IEEE. https://doi.org/10.1109/ICICCS56967.2023.10142531.

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

[22] Han, G., Huang, S., Ma, J., He, Y., & Chang, S. F. (2022, June). Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 1, pp. 780-789). https://doi.org/10.1609/aaai.v36i1.19959.

[23] Yang, H., Liu, P., Hu, Y., & Fu, J. (2021). Research on underwater object recognition based on YOLOv3. Microsystem Technologies, 27, 1837-1844. https://doi.org/10.1007/s00542-019-04694-8.

[24] Shen, L., Tao, H., Ni, Y., Wang, Y., & Stojanovic, V. (2023). Improved YOLOv3 model with feature map cropping for multi-scale road object detection. Measurement Science and Technology, 34(4), 045406. http://dx.doi.org/10.1088/1361-6501/acb075.

[25] Bosse, S., & Kasundra, P. (2022). Robust Underwater Image Classification Using Image Segmentation, CNN, and Dynamic ROI Approximation. Engineering Proceedings, 27(1), 82. https://doi.org/10.3390/ecsa-9-13218.

[26] Chen, Z., Wang, Y., Tian, W., Liu, J., Zhou, Y., & Shen, J. (2022). Underwater sonar image segmentation combining pixel-level and region-level information. Computers and Electrical Engineering, 100, 107853. https://doi.org/10.1016/j.compeleceng.2022.107853.

[27] Wang, J., He, X., Shao, F., Lu, G., Hu, R., & Jiang, Q. (2022). Semantic segmentation method of underwater images based on encoder-decoder architecture. Plos one, 17(8), e0272666. https://doi.org/10.1371/journal.pone.0272666.

[28] Liu Z, Tong L, Chen L, Zhou F, Jiang Z, Zhang Q, et al. Canet: Context aware network for brain glioma segmentation. 2021;40(7):1763-77. https://doi.org/10.1109/tmi.2021.3065918.

[29] Alavianmehr, M. A., Helfroush, M. S., Danyali, H., & Tashk, A. (2023). Butterfly network: a convolutional neural network with a new architecture for multi-scale semantic segmentation of pedestrians. Journal of real-time image processing, 20(1), 9. https://doi.org/10.1007/s11554-023-01273-z.

[30] Dakhil, R. A., & Khayeat, A. R. H. (2022). Review On Deep Learning Technique For Underwater Object Detection. arXiv preprint arXiv:2209.10151. https://doi.org/10.48550/arXiv.2209.10151.

[31] Islam, M. J., Enan, S. S., Luo, P., & Sattar, J. (2020, May). Underwater image super-resolution using deep residual multipliers. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 900-906). IEEE. https://doi.org/10.1109/ICRA40945.2020.9197213.

[32] Islam MJ, Luo P, Sattar JJapa. Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. 2020. https://doi.org/10.48550/arXiv.2002.01155.

[33] Marine Life Encyclopedia 2001 [cited 2023. Available from: https://oceana.org/marine-life/.

[34] Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, et al., editors. Esrgan: Enhanced super-resolution generative adversarial networks. Proceedings of the European conference on computer vision (ECCV) workshops; 2018. https://doi.org/10.1007/978-3-030-11021-5_5.

[35] Aghelan, A. (2022). Underwater Images Super-Resolution Using Generative Adversarial Network-based Model. arXiv preprint arXiv:2211.03550. https://doi.org/10.48550/arXiv.2211.03550.

[36] Rakotonirina NC, Rasoanaivo A, editors. ESRGAN+: Further improving enhanced super-resolution generative adversarial network. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020: IEEE. https://doi.org/10.48550/arXiv.2001.08073.

[37] 33. Wang, H., Zhong, G., Sun, J., Chen, Y., Zhao, Y., Li, S., & Wang, D. (2023). Simultaneous restoration and super-resolution GAN for underwater image enhancement. Frontiers in Marine Science, 10, 1162295. https://doi.org/10.3389/fmars.2023.1162295.

[38] 34. Zhang Z, Sabuncu MJAinips. Generalized cross-entropy loss for training deep neural networks with noisy labels. 2018;31. https://doi.org/10.48550/arXiv.1805.07836.

[39] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al., editors. {TensorFlow}: a system for {Large-Scale} machine learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16); 2016.

[40] Kingma DP, Ba JJapa. Adam: A method for stochastic optimization. 2014. https://doi.org/10.48550/arXiv.1412.6980.

[41] Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., ... & Sattar, J. (2020, October). Semantic segmentation of underwater imagery: Dataset and benchmark. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1769-1776). IEEE. https://doi.org/10.1109/IROS45743.2020.9340821.

Radhwan Adnan Dakhil, is a rising computer science scholar born in Karbala, Iraq, in 1996. Currently pursuing a master's degree at the University of Kerbala, she specializes in Information Science, Artificial Intelligence, and Data Mining. With expertise in classification, machine learning, computer vision, and object detection, Radhwan is driven by a passion for leveraging technology to address complex challenges. Fluent in both Arabic and English, she fosters global connections in the field of computer science, poised to make impactful contributions to research and innovation.

Ali Retha Hasoon Khayeat, is a renowned engineering expert with a Doctor of Engineering degree. Serving as a lecturer at the University of Kerbala in the United Kingdom, she excels in fields such as pattern recognition, computer vision, image processing, and machine learning. Her proficiency extends to both English and Arabic, enabling her to engage in global research and collaboration, making her a valuable contributor to the world of engineering and academia.