Underwater Semantic Segmentation using Multi-scale Feature Extraction

A Minor B. Tech Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Devika Singh
(210101036)

under the guidance of

Dr. Arijit Sur

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Underwater Semantic Segmentation using Multi-scale Feature Extraction**" is a bonafide work of **Devika Singh** (**Roll No. 210101036**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. Arijit Sur**

Professor,

Nov, 2024

Department of Computer Science & Engineering,

Guwahati.

Indian Institute of Technology Guwahati, Assam.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, **Professor Dr. Arijit Sur**, and my mentor, **Alik Pramanik**, for their invaluable support and continuous guidance throughout this project. Their expertise and encouragement have greatly contributed to my learning and the successful completion of my BTP work. I am truly thankful for their valuable time and the constant efforts they invested in assisting me during this journey.

Abstract

This BTP enhances the SUIM-Net model for real-time semantic segmentation of underwater imagery by introducing a Parallel Dilation Convolution Block and integrating the Convolutional Block Attention Module (CBAM) within the model's skip connections. These modifications improve the model's ability to handle multi-scale features and focus on salient regions, thereby enhancing segmentation accuracy in complex underwater scenes. Additionally, pixel shuffle pooling replaces traditional methods to preserve high-resolution details. An analysis of the original SUIM-Net$_{VGG}$ underscores the improvements and identifies persistent challenges, guiding future enhancements. This work advances underwater robotic vision by boosting the efficiency and adaptability of SUIM-Net for environmental monitoring.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Semantic segmentation is the process of categorizing each pixel in an image into predefined object classes. Unlike object detection, which identifies and localizes objects with bounding boxes, semantic segmentation provides a fine-grained understanding of a scene by assigning a label to every pixel.

For example, in an underwater image, it can distinguish between pixels representing fish, plants, reefs, and the sea-floor. Semantic segmentation is crucial for a wide range of applications. In robotics, it enables detailed scene understanding, facilitating autonomous navigation and decision-making. Robots can infer spatial relationships, identify obstacles, and interact effectively with their environment.

In domains like underwater exploration, semantic segmentation aids in mapping and monitoring ecosystems, identifying objects of interest (e.g., wrecks or coral reefs), and assisting human-robot collaboration. Its pixel-level precision is especially important in tasks like saliency prediction, where focusing on specific regions enhances operational accuracy. As a foundation for advanced tasks like visual question answering or spatio-temporal attention modeling, semantic segmentation is a cornerstone of visual perception systems.

1.1 Challenges

The following are key challenges we will tackle in enhancing SUIM-Net model:

1.1.1 Lack of Feature Prioritization

The original SUIM-Net architecture relies heavily on residual skip connections, employed to help recover spatial details lost during down-sampling in the encoder. These connections carry feature maps from the encoder directly to corresponding layers in the decoder, facilitating precise localization and detailed segmentation. However, they treat all features equally without determining which feature is crucial for the task. Thus non-essential features may interfere with or dilute the impact of more important features.

1.1.2 Handling Varying Object Sizes

Underwater environments include objects of vastly different scales, such as small fish and large shipwrecks. The fixed-scale feature extraction in the original SUIM-Net model could not adapt to this variability effectively. As a result, smaller objects were often missed, while larger objects lacked detailed segmentation. This variability necessitated a method to dynamically process features at multiple scales to ensure robust segmentation for all object types.

1.1.3 Loss of Spatial Information

Underwater scenes often exhibit fine-grained textures and subtle differences in object boundaries that require higher resolution features. Both max and average pooling involve summarizing or down-sampling information to reduce the spatial dimensions of feature maps. This leads to a significant loss of fine spatial details, which is problematic for tasks requiring high-resolution outputs or precise localization, such as image segmentation.

1.2 Problem Statement

The SUIM-Net model, introduced in [IEX+20], demonstrated efficient semantic segmentation of underwater imagery. However, it faced limitations in capturing multi-scale features, leading to challenges in segmenting objects of varying sizes and complexities in diverse underwater environments. This highlighted the need for an enhancement to improve its segmentation accuracy without compromising computational efficiency.

We enhanced SUIM-Net by incorporating Parallel Dilation Convolution Encoder Blocks, improving its ability to handle varying object sizes while maintaining real-time efficiency. We also used Convolutional Block Attention Module (CBAM)[WPLK18] in Skip Connections for selective feature concatenation to significantly boost the quality of the segmentation by reducing the background noise and focusing on salient objects. This modification and its impact on segmentation accuracy will be detailed further in this report.

1.3 Our contributions

To address this, we propose incorporating a Parallel Diverse Dilation Convolution Block into the encoder. This block utilizes parallel dilated convolutions, each with varying dilation rates, enabling the model to capture contextual details across different spatial extents efficiently. The parallel structure of these convolutions is computationally more efficient than the serial convolution layers used in SUIM-Net's encoder blocks, as it allows simultaneous processing of features at various scales.

Additionally, the integration of the Convolutional Block Attention Module (CBAM) refines these features, ensuring that only the most relevant details are emphasized. Pixel shuffle pooling is also employed instead of traditional methods, further enhancing the model's efficiency and preserving more spatial information, critical for real-time applications in underwater robotic systems.

- **Lack of Feature Prioritization:** We integrated CBAM to selectively emphasize crucial features, minimizing the impact of non-essential ones.
- **Handling Varying Object Sizes:** We implement parallel dilated convolutions to effectively capture features at varying spatial extents, enhancing object detection across sizes.
- **Loss of Spatial Information:** We use pixel shuffle pooling to retain finer spatial details and high-resolution features essential for precise segmentation.

These improvements enhance segmentation accuracy while ensuring the model remains suitable for real-time applications. Details of this modification and its impact on performance will be elaborated further in this report.

1.4 Organization of the Report

This report is structured to provide a comprehensive overview of the proposed enhancement to the SUIM-Net architecture. Chapter 1 introduces the problem, highlighting challenges in underwater semantic segmentation and the motivation behind our work. Chapter 2 reviews related works in semantic segmentation techniques, focusing on the use of Convolutional Block Attention Module (CBAM), semantic segmentation, and their applications in enhancing underwater imagery analysis. Chapter 3 details the proposed enhanced SUIM-Net. Chapter 4 explains the experimental setup, dataset, and evaluation metrics used to validate our model. Chapter 5 presents the results and analysis of the proposed method, while Chapter 6 concludes the report and outlines potential future directions.

Chapter 2

Related Works

2.1 Semantic Segmentation Techniques

Fully Convolutional Networks (FCNs) [LSD15] introduced end-to-end learning for semantic segmentation, enabling pixel-wise predictions. This foundation inspired encoder-decoder architectures like U-Net [RFB15] and SegNet [BKC15], which effectively handle segmentation tasks by utilizing skip connections. These advancements directly influence the design of SUIM-Net's encoder-decoder structure.

2.2 SUIM Dataset and SUIM-Net model[IEX$^+$20]

The SUIM[IEX$^+$20] dataset is a large-scale, annotated underwater image collection designed for semantic segmentation, containing over 1,500 images with pixel-level labels across eight object categories, including fish, reefs, and robots.

SUIM-Net[IEX$^+$20] is a fully-convolutional encoder-decoder model optimized for underwater imagery, balancing competitive segmentation accuracy with fast inference for real-time robotic applications.

The SUIM dataset and SUIM-Net model[IEX$^+$20] are critically important for advancing underwater robotic vision. The SUIM dataset and SUIM-Net are designed to advance

semantic segmentation in underwater environments, addressing challenges like optical distortions and domain-specific object categories. By providing a large-scale annotated dataset and an efficient segmentation model, they enable precise pixel-level scene understanding critical for underwater exploration, robotic navigation, and human-robot collaboration.

2.3 Convolutional Block Attention Module (CBAM)[WPLK18]

CBAM[WPLK18] is an attention mechanism for deep neural networks that sequentially infers attention maps along channel and spatial dimensions, enhancing feature representation. It highlights important features by focusing on informative regions and channels within the image. A key use case of CBAM is in underwater image segmentation, where it enhances the detection and delineation of submerged objects by emphasizing crucial features. This improves the clarity and accuracy of segmenting objects like marine species and underwater structures, essential for tasks in marine biology research and underwater exploration.

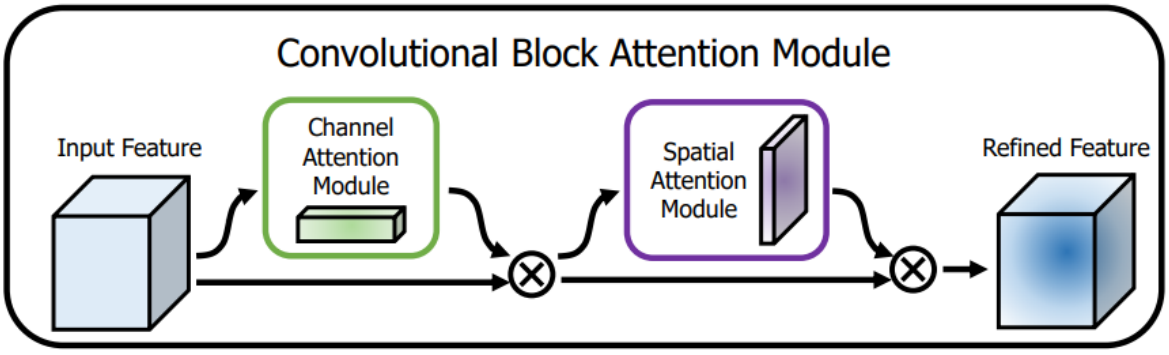


Fig. 2.1: CBAM module (referenced from [WPLK18]

2.4 Underwater Semantic Segmentation

Several works focus on underwater segmentation in applications like coral reef classification [BEK+12] and fish detection [RSS+15]. However, these are often limited to binary

annotations or small datasets. The SUIM dataset extends these efforts with comprehensive annotations for diverse underwater objects.

2.5 Few-Shot Segmentation and Semantic Segmentation for Underwater Imagery

This paper[KSM+23] introduces a novel underwater image dataset and explores few-shot and semantic segmentation using attention-guided deep neural networks. It innovatively addresses the scarcity of underwater datasets suitable for complex segmentation tasks and demonstrates significant improvements in segmentation accuracy. The approach potentially transforms underwater imagery analysis, making it invaluable for marine biology and robotics. The dataset and code are openly available, fostering further research in this critical area.

2.6 Computational Efficiency

The need for efficient real-time models has led to lightweight architectures like MobileNet [HZC+17]. SUIM-Net builds on this principle, balancing performance and speed, making it suitable for visually guided underwater robots.

2.7 Conclusion

This chapter outlined the foundational advancements in semantic segmentation, the importance of multi-scale feature extraction, and existing works in underwater segmentation. These insights directly inform the design of SUIM-Net, highlighting its improvements in accuracy and efficiency for underwater robotic applications.

Chapter 3

Proposed Work

3.1 Models

The proposed work builds upon the SUIM-Net architecture, incorporating a multi-scale feature extraction mechanism to improve its segmentation performance on underwater imagery. This modification addresses challenges in capturing diverse object scales and complex scene features, enhancing the model's generalization and accuracy without compromising computational efficiency.

3.2 Architecture Details

he enhanced SUIM-Net maintains its encoder-decoder structure with skip connections but now incorporates parallel dilation convolutions multi-scale feature extraction.

The main components are as follows:

3.2.1 Parallel Dilation Convolution Block.

This block configuration leverages parallel dilated convolutions and pixel shuffle pooling to efficiently handle and process multi-scale feature maps, ensuring comprehensive and effective feature extraction across varying spatial extents.

This block takes a feature map as an input. The final output is also a feature map after pixel shuffle pooling, provides a comprehensive, multi-scale feature map that is ready for further processing in the network. Each block in the four serial encoders are equipped with an increasing number of filters: 64, 128, 256, and 512 filters, allowing for more complex feature extraction as the spatial extents increase Components of the Block are:

- **Parallel Dilated Convolutions:**
  - **Convolution 1:** 3x3 kernel, dilation factor of 1. Captures fine details.
  - **Convolution 2:** 3x3 kernel, dilation factor of 2. Expands the receptive field to capture medium-scale features.
  - **Convolution 3:** 3x3 kernel, dilation factor of 4. Further expands the receptive field for large-scale feature capture.
- **Feature Fusion via Concatenation:** Outputs from each of the three dilated convolution layers are concatenated along the channel dimension. This method preserves all feature information from different scales, enriching the feature representation with details from varying perspectives and scales.
- **Pixel Shuffle Pooling:** Takes the concatenated feature map and applies a pixel shuffle operation to perform downsampling. This advanced pooling technique rearranges elements to reduce spatial dimensions while aiming to preserve more textual and structural information than traditional pooling methods.

3.2.2 Integration with SUIM-Net

The integration of enhancements into SUIM-Net$_{\text{VGG}}$ includes significant architectural modifications designed to improve segmentation accuracy and feature representation:

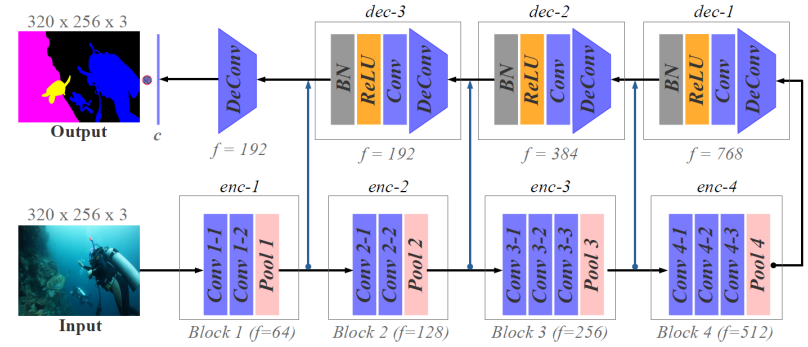
- The serial encoder blocks in SUIM-Net$_{\text{VGG}}$ have been replaced by a **Parallel Dilation Convolution Block**. Each encoder block uses parallel processing with varying

dilation rates to capture a broader range of contextual details, more efficiently.

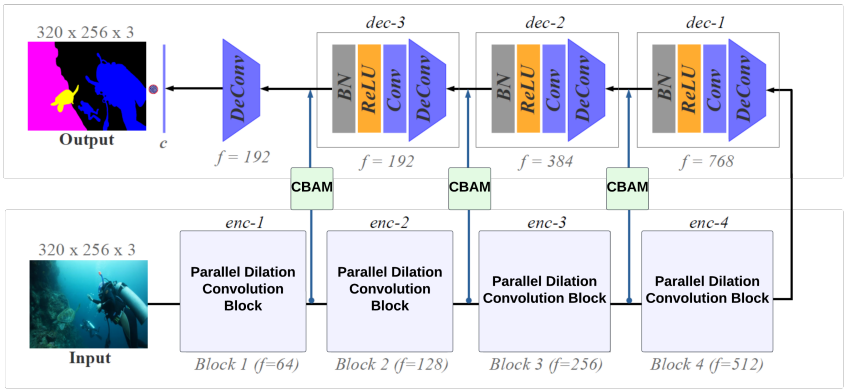
- The skip connections in SUIM-Net$_{\text{VGG}}$ have been augmented with the **Convolutional Block Attention Module (CBAM)**[WPLK18], which refines the features passed through these connections. By emphasizing relevant features and suppressing less useful ones, CBAM enhances the model's ability to distinguish between important elements in complex scenes. For example, CBAM helps the network to better focus on and segment a small, distant diver in the background, ensuring that even subtle features are not overshadowed by dominant foreground elements.

3.3 Conclusion

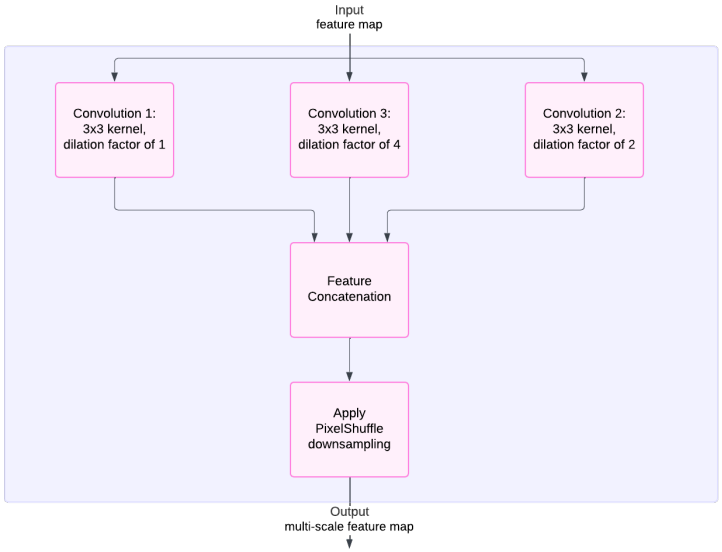
In this chapter, we proposed enhancements to the SUIM-Net$_{\text{VGG}}$ architecture, introducing a Parallel Dilation Convolution Block and integrating the Convolutional Block Attention Module (CBAM)[WPLK18] with skip connections. These improvements allow for more efficient multi-scale feature extraction and enhance feature refinement, enabling the network to effectively highlight subtle yet critical features, such as a distant diver, improving segmentation accuracy and computational efficiency in complex underwater environments. Additionally, the incorporation of pixel shuffle pooling enhances the model's ability to maintain high-resolution details during downsampling.



(a)



(b)



(c)

Fig. 3.1: Overview of the images: (a) *SUIM-Net$_{VGG}$* end-to-end architecture {referenced from [IEX+20]}, (b) Enhanced *SUIM-Net$_{VGG}$*: Integrating Parallel Dilation Convolution Encoder Blocks with dilations rates = 1, 2 and 4, and CBAM in Skip Connections for Improved Underwater Segmentation, and (c) Internal architecture of the Parallel Dilation Convolution Block.

Chapter 4

Experimental Setup

4.1 Setup

We used Google Colab to run the project code, for its cloud-based computational resources.

4.2 Dataset

We use the SUIM dataset [IEX+20] in our work to evaluate and enhance semantic segmentation techniques for underwater imagery.

Object Category	**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

Table 4.1: The object categories and their associated color codes for pixel annotations in the SUIM dataset, [IEX+20], which our work references.

4.3 Evaluation Metrics

In this work, we use the same evaluation metrics as those in the SUIM paper [IEX+20] to ensure consistent benchmarking of segmentation performance. These metrics include:

- **F-Score (Dice Coefficient):** The F-Score evaluates the balance between precision and recall for pixel-level segmentation. It is calculated as:

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

  where P is precision, and R is recall.

- **Mean Intersection over Union (mIoU):** This metric evaluates the extent of overlap between the predicted regions and the ground-truth regions. It is defined as:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

  The mIoU averages this score across all object categories.

These metrics are well-suited for evaluating region similarity and contour accuracy in semantic segmentation, particularly in challenging underwater environments, as demonstrated in the SUIM dataset.

Chapter 5

Conclusion

In this BTP, we enhanced the SUIM-Net$_{\text{VGG}}$ architecture by incorporating a Parallel Dilation Convolution Block and integrating CBAM into its skip connections to refine multi-scale feature extraction for underwater imagery. Additionally, pixel shuffle pooling was implemented to preserve high-resolution details crucial for precise segmentation.

We trained the original SUIM-Net$_{\text{VGG}}$ model on SUIM dataset, to understand its operational mechanics and to identify the challenges it faces in segmenting diverse underwater scenes. These insights have been instrumental in guiding our modifications and will inform continued efforts to address the complexities of underwater imagery segmentation.

We conducted tests on the modified SUIM-Net$_{\text{VGG}}$ to evaluate its performance and analyze its capabilities.

Future work will focus on addressing the issues faced by the enhanced SUIM-Net$_{\text{VGG}}$ model to enhance the model's efficacy further.

References

[BEK+12] Oscar Beijbom, Peter J Edmunds, David I Kline, B Greg Mitchell, and David Kriegman. Automated annotation of coral reef survey images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1170–1177, 2012.

[BKC15] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[HZC+17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[IEX+20] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. *arXiv preprint arXiv:2004.01241*, 2020.

[KSM+23] Imran Kabir, Shubham Shaurya, Vijayalaxmi Maigur, et al. Few-shot segmentation and semantic segmentation for underwater imagery. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. https://github.com/Imran2205/uwsnet.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[RSS+15] Mohammad Ravanbakhsh, Mark R Shortis, Faisal Shafait, Ajmal Mian, Euan S Harvey, and James W Seager. Automated fish detection in underwater images using shape-based level sets. *Photogrammetric Record*, 30(149):46–62, 2015.

[WPLK18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. *arXiv preprint arXiv:1807.06521*, 2018.