Deep Learning Models for Underwater Image Enhancement and Super-Resolution

Thesis submitted to the
Indian Institute of Technology Guwahati
for the award of the degree

of

Doctor of Philosophy

in

Computer Science and Engineering

Submitted by

Alik Pramanick

Under the guidance of

Prof. Arijit Sur

and

Prof. V. Vijaya Saradhi

Department of Computer Science and Engineering

Indian Institute of Technology Guwahati

May 17, 2026

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Prof. Arijit Sur
and
Prof. V. Vijaya Saradhi
Department of Computer Science and Engineering
IIT Guwahati
Email : arijit@iitg.ac.in and saradhi@iitg.ac.in

Certificate

This is to certify that this thesis entitled, **"Deep Learning Models for Underwater Image Enhancement and Super-Resolution"**, being submitted by **Alik Pramanick**, to the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, for partial fulfillment of the award of the degree of Doctor of Philosophy, is a bonafide work carried out by him under our supervision and guidance. The thesis, in our opinion, is worthy of consideration for award of the degree of Doctor of Philosophy in accordance with the regulations of the institute. To the best of our knowledge, it has not been submitted elsewhere for the award of the degree.

.....
Prof. Arijit Sur
Professor

.....
Prof. V. Vijaya Saradhi
Professor

Date: May 17, 2026
Place: Guwahati

Declaration

I certify that:

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisor.
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding, and I take complete responsibility if any complaint arises.
- I am fully aware that my thesis supervisor is not in a position to check for any possible instance of plagiarism within this submitted work.

Date: May 17, 2026
Place: Guwahati

(Alik Pramanick)

Dedicated to

My Family & Friends

For their unconditional love, blessings, and constant inspiration

Acknowledgements

My Ph.D. journey has been a meaningful and inspiring adventure, shaped by the patience, encouragement, and wisdom of many people. I extend my heartfelt gratitude to everyone who supported me throughout the preparation of this doctoral thesis.

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Prof. Arijit Sur and Prof. V. Vijaya Saradhi, for their unwavering guidance, encouragement, and patience throughout this journey. I am especially grateful to Prof. Arijit Sur for helping me stay focused and disciplined in my research. His clarity of thought and emphasis on rigorous presentation have significantly shaped my research. I am equally indebted to Prof. V. Vijaya Saradhi for his consistent support and the time he generously devoted to guiding me at every stage. His understanding and flexibility helped me handle challenges with confidence. I truly appreciate the freedom they gave me to explore ideas while always being available to offer direction and motivation whenever needed.

I would also like to express my sincere gratitude to the members of my Doctoral Committee - Prof. Sanasam Ranbir Singh, Prof. Pinaki Mitra, and Prof. Santosha K. Dwivedy for their insightful comments and suggestions, which greatly contributed to improving the quality of my work. I am further thankful to the anonymous reviewers of my research for their thoughtful and critical feedback, which helped me enhance the quality of my work.

I would like to express my sincere gratitude to Prof. Tamarapalli Venkatesh, Head of the Department of Computer Science and Engineering at IIT Guwahati, for providing access to the department's facilities and resources. I am thankful to all the faculty members, staff, and security personnel for their constant assistance and support. I also extend my appreciation to the staff of the Academic Affairs office for their support in processing my applications and grant requests. I gratefully acknowledge the MHRD, Government of India, for the financial support provided throughout my Ph.D. years, without which this research would not have been possible.

I am grateful to my seniors, Anirban Lekharu, Sandipan Sarma, Avinash Kumar Chouhan, and Prasen Kumar Sharma, for their initial support and guidance. I

would also like to extend my sincere thanks to the BTP and MTP students who worked with me, Bhaskar, Tanvish, Dhruvil, Utsav, Lovish, Jaya Krishna, Priyanshu, Soumajit, and many others, for their unconditional support. I extend my appreciation to my fellow research scholars and friends, Laxita, Sonal, Akshay, Adithya, Soumya, Mishra, Sreyas, Rajeswari, Soumyadeep, Nilmadhab, Suklav, Utkarsh, Suvarthi, Shubhradeep, Prachuryya, Rohit, Swati, Praveen, Nilotpola, Ajanta, Saurav Da, Shifali, and Vivekananda, for making my Ph.D. journey both memorable and enjoyable. Their encouragement, support, and willingness to help me navigate academic and personal challenges have been invaluable. Our discussions and exchanges of ideas have greatly enriched my research, and their companionship has been an essential part of this experience. I am deeply grateful to my hostel friends Partha, Santanu, Adri, Tanay, Aritra, Anit, Balaram, Arin da, Anjishnu da, Sanket da, Surojit da, and Anindya da for their unwavering support, camaraderie, and the warmth they brought into my everyday life. I would like to thank Bivakar and Bikram for their wonderful home-cooked meals, which provided comfort and joy during the demanding phases of my research. I am especially thankful to Dr. Sukanta Bhattacharjee for his kindness and hospitality, as well as for sharing memorable moments on the badminton court and cricket field. I am equally thankful to Ujjwal da and the morning football group, whose enthusiasm made every day more lively and refreshing. Additionally, I would like to acknowledge the IIT Guwahati tennis ball cricket groups for the fun-filled matches and the spirited moments we shared, which added joy and relaxation amidst the rigors of research.

I am profoundly grateful to God and to my family, including my brother Ayan, parents, and grandmother, for their unconditional love, support, care, warmth, and profound encouragement throughout the years. My heartfelt gratitude goes especially to my parents, whose trust and confidence in me made this journey meaningful, secure, and fulfilling. I fall short of words to express how much their support has meant to me.

Abstract

Recent advances in underwater exploration have attracted significant attention due to their wide-ranging applications in underwater navigation, marine warfare, oceanography, and marine scene understanding or measurements. In this case, computer vision plays a pivotal role because high-quality images are necessary for terrestrial workstations, which are captured by underwater surveillance systems and Autonomous Underwater Vehicle (AUV) through a camera. However, underwater image acquisition is inherently challenging. Images often suffer from low contrast, haziness, and blurriness due to the presence of suspended particles, plankton, and variations in light. They also suffer from strong color distortion, which causes them to appear blue or green due to the way light is absorbed underwater. These degradations make underwater image enhancement a critical research problem. Even after enhancement, many small objects, such as tiny fish, cables, coral tips, or trash, can still be hard to see clearly at the original resolution. This negatively impacts downstream tasks, such as object classification, detection, and segmentation, making them less accurate. To solve this, Super-resolution (SR) is necessary to recover sharper textures and more detailed object boundaries, making small or distant objects more visible in high-resolution images.

Nowadays, researchers are focusing their work on Deep Learning (DL)-based methods due to their superior performance compared to traditional techniques. The existing literature clearly demonstrates that learning-based methods for enhancement and SR excel not only in performance but are also suitable for deployment in resource-constrained or real-time environments. With this motivation, this thesis presents efficient deep learning-based models for two specific tasks: (a) Underwater Image Enhancement (UIE) and (b) Underwater Image Super-resolution (UISR).

To enhance underwater images, the first chapter presents two studies that focus on learning features based on color channels. These methods consider how different wavelengths of light are affected differently, which causes colors to appear unevenly. The first work proposes a lightweight Convolutional Neural Network (CNN)-based multi-stage model called Lit-Net, which focuses on multi-resolution and multi-scale image analysis to enhance underwater images. Knowing that the underwater images are mainly bluish and greenish, removing only the undesired degraded content is crucial while preserving all the essential spatial information, including true edges and texture. This work utilizes different sizes of receptive fields for each color channel: R (3×3), G (5×5), and B (7×7), enabling the model to analyze images at various resolutions while preserving important details. It also processes the entire RGB image with a 1×1 convolution to enhance global effects and maintain color relationships, helping to connect with earlier layers. Moreover, the multi-scale strategy retains important information while improving efficiency. Extensive experimenta-

tion on publicly available datasets suggests that LitNet outperforms recent state-of-the-art (SOTA) methods, with significant improvements in both qualitative and quantitative measures. In contrast, the second work, X-CAUNET, employs multiple kernel sizes (3×3, 5×5, and 7×7) for each color channel, allowing the model to simultaneously capture fine details (small kernels) and broader contextual structures (larger kernels) within each channel. This avoids restricting a channel to a single receptive field that may not be optimal for all underwater conditions. Then, it performs cross-channel message passing to enhance correlations between channels, resulting in improved robustness against depth variations, better color consistency, and more reliable structural preservation in UIE.

The second contributory chapter extends channel-aware enhancement by presenting two additional methods: a) dual-color space feature fusion and b) spatial-frequency domain modeling to better handle the complex color and structural degradations in underwater images. The first work of this second contributory chapter proposes UEnhancer, which incorporates a dual-color fusion-based hierarchical encoder-decoder architecture, leveraging the capabilities of a Transformer to enhance images effectively. To maintain the local structure, contrast in high-frequency regions, and retain intricate detail information, the proposed model aims to minimize a multi-scale frequency-based SSIM loss. A comprehensive set of experiments reveals that the proposed UEnhancer outperforms other SOTA methods on several publicly accessible underwater image enhancement datasets. The second work of this chapter, D2Mamba, further improves underwater enhancement by jointly modeling spatial and frequency cues using a state space model guided by an A*-based traversal strategy and a physics-inspired Geodesic Information-Field Heuristic. This adaptive dual-domain processing, along with a Spectral Wasserstein Attenuation Loss (SWAL), enforces distributional alignment in the spectral domain, leading to better color recovery and accurate reconstruction of fine details.

While the earlier models improved enhancement performance for specific types of degradation, real underwater environments frequently suffer from multiple degradations simultaneously. Models developed for one type of degradation usually struggle in diverse conditions, which limits their efficacy in real-world situations. To address this problem, the third contributory chapter presents a unified multi-degradation enhancement model that can simultaneously handle various types of underwater degradation. First, the model generates various degraded versions of the input image, including haze, blur, low contrast, and color distortion, through domain translation (DT). Then, the original degraded images and the translated ones are input into a multi-degraded enhancement model. This model uses several encoders with an intermediate feature-swapping mechanism and a single decoder. Extensive experiments have demonstrated that incorporating DT significantly improves the model's performance on cross-dataset evaluation.

Moving beyond ocean conditions, in riverine environments, images are severely degraded

due to the presence of clay, silt, sand, and other particles. To tackle this issue, the fourth contributory chapter introduces River-GEM, which includes (1) a dataset generation of highly degraded muddy water images using a principal component analysis-based fusion technique and (2) a novel enhancement model utilizing the depthmap guidance to learn complex contextual details of muddy water images. The proposed model captures the intricate details of the objects in degraded images by analyzing feature correlation at multiple scales. This dataset and enhancement model mark a significant step forward in this area of research.

Following the focus on enhancement, the fifth contributory chapter shifts to UISR. Most recent research has primarily analyzed images in the spatial domain for UISR, ignoring the fact that frequency domain information can help super-resolve complex high-frequency image regions. Keeping this in mind, the first work of this fifth contributory chapter proposes a model that amalgamates the Discrete Wavelet Transform (DWT) and the effective utilization of attention modules to enhance the performance of UISR. Further, the Difference of Gaussian (DoG) loss is employed to preserve the quality and enhance the sharpness of edges in an SR image, as the edges become blurry when an Low-resolution (LR) image is super-resolved. In contrast, the second work of this chapter proposes Efficient-USR, a dual-domain (spatial and Fast Fourier Transform (FFT)) information-based lightweight model for UISR. In Efficient-USR, images are analyzed at different scales by incorporating two key components: a) a spatial-frequency interaction block to capture both the global and local context by operating spatial-channel cross-attention on spatial-frequency features, and b) a prompt-guided cross-attention block to efficiently encode degradation information in the prompt component and perform feature fusion of multiple scales via cross-attention. Both methods reveal that incorporating spatial and frequency details, whether through localized wavelet decomposition or global Fourier analysis, is critical for achieving high-quality and efficient underwater image super-resolution.

Finally, the thesis concludes by summarizing its significant contributions and proposing some relevant future research directions.

Contents

Certificate iv

Declaration vi

Dedication viii

Acknowledgement x

Abstract xii

Contents xvi

List of Figures xxii

List of Tables xxviii

List of Algorithms xxxii

List of Symbols xxxiv

List of Acronyms xxxvii

1 Introduction 1

1.1 Underwater Image Enhancement 4

1.2 Underwater Image Super-Resolution 4

1.3 Applications 5

1.4 Motivation of the Research Work 6

1.5 Objectives of the Thesis 7

1.6 Contributions of the Thesis 7

1.6.1 Color Channel-Aware Feature Learning for Ocean-water Image Enhancement 8

1.6.2 Dual Feature Transformations for Ocean-water Image Enhancement . 8

1.6.3 Multi-Degradation Learning Framework for Oceanic Image Enhancement . . . 9
1.6.4 Generating and Enhancing River-water Images . . . 9
1.6.5 Dual-Domain Attention Model for Underwater Image Super-Resolution 10
1.7 Organization of the Thesis . . . 11

2 Research Background and Literature Survey 13
2.1 Convolutional Neural Network . . . 14
2.1.1 Standard Convolution . . . 14
2.1.2 Depthwise Separable Convolution . . . 14
2.1.3 Dilated Convolution . . . 15
2.1.4 Deformable Convolution . . . 15
2.1.5 Strip Convolution . . . 15
2.2 Transformer . . . 15
2.3 State Space Model . . . 16
2.4 Activation Functions . . . 16
2.5 Pooling Layer . . . 18
2.6 2D Discrete Wavelet Transform . . . 19
2.7 2D Fast Fourier Transform . . . 21
2.8 Difference of Gaussian. . . . 21
2.9 Literature Survey . . . 23
2.9.1 Underwater Image Enhancement . . . 23
2.9.1.1 Physical Model-based methods . . . 23
2.9.1.2 Physical Model-free methods . . . 24
2.9.1.3 Deep Learning-based methods . . . 24
2.9.2 Underwater Image Super-Resolution . . . 26
2.9.2.1 Interpolation-based methods . . . 26
2.9.2.2 Deep Learning-based methods . . . 27
2.10 Limitations of existing works . . . 28
2.11 Evaluation Metrics . . . 32
2.11.1 Full-reference Metrics . . . 32
2.11.2 Reference-less metrics . . . 35
2.12 Datasets . . . 37
2.13 Summary . . . 38

3 Color Channel-aware Feature Learning for Ocean-water Image Enhancement 39

3.1 Lit-Net: Harnessing Multi-resolution and Multi-scale Attention for Underwater Image Enhancement . . . 40
3.1.1 Proposed Method . . . 42
3.1.1.1 Network Architecture . . . 42
3.1.1.2 Multi-resolution Attention Network . . . 43
3.1.1.3 Multi-scale Attention Network . . . 45
3.1.1.4 Image Reconstruction . . . 47
3.1.1.5 Loss Function . . . 47
3.1.2 Experiments . . . 49
3.1.2.1 Experimental Setup . . . 50
3.1.2.2 Comparisons with State-of-the-Arts . . . 50
3.1.3 Ablation Study . . . 54
3.1.3.1 Impact on High level Vision Applications . . . 57
3.1.3.2 Failure Case . . . 60
3.2 X-CAUNET: Cross-color Channel Attention with Underwater Image-enhancing Transformer . . . 61
3.2.1 Proposed Model . . . 61
3.2.1.1 Cross-color channel attention transformer (C^3A) . . . 62
3.2.1.2 Aggregating global context . . . 64
3.2.1.3 Loss functions . . . 64
3.2.2 Experiments . . . 66
3.3 Summary . . . 67

4 Dual Feature Transformations for Ocean-water Image Enhancement 71
4.1 UEnhancer: Spatially Enriched Transformer with Dual-color Information for Underwater Image Enhancement . . . 72
4.1.1 Proposed Method . . . 73
4.1.1.1 Overall Pipeline . . . 74
4.1.1.2 Transformer Block . . . 75
4.1.1.3 Dual-color Interaction (DCI) block . . . 77
4.1.1.4 Latent Space Fusion with Convolution Attention . . . 79
4.1.1.5 Loss Function . . . 79
4.1.2 Experiments . . . 80
4.1.2.1 Experimental Settings . . . 80
4.1.2.2 Comparisons with SOTA schemes . . . 82
4.1.2.3 Ablation Study . . . 85
4.1.2.4 Impact on Various Applications . . . 90
4.1.2.5 Generalization performance of the proposed method . . . 91

4.2 D2Mamba: Dual Domain Guided Informed Search in State Space Model for Underwater Image Enhancement . . . 92
4.2.1 Proposed Method . . . 95
4.2.1.1 Overall Model Integration . . . 96
4.2.1.2 Dual-Stream Encoder (DSE) . . . 97
4.2.1.3 Enhanced A* Pathfinding with Affinity-Based Edge Costs . . . 98
4.2.1.4 Geodesic Information-Field Heuristic (GIFH) . . . 98
4.2.1.5 Gated Synthesis Decoder (GSD) . . . 100
4.2.1.6 Training Losses . . . 100
4.2.2 Experiments . . . 102
4.2.2.1 Implementation Details . . . 102
4.2.2.2 Datasets . . . 103
4.2.2.3 Comparison Methods . . . 103
4.2.2.4 Evaluation Metrics . . . 103
4.2.2.5 Quantitative Evaluation . . . 104
4.2.2.6 Qualitative Evaluation . . . 104
4.2.2.7 Ablation Study . . . 105
4.3 Summary . . . 106

5 Multi-degradation Learning Model for Ocean-water Image Enhancement 109
5.1 Proposed Method . . . 112
5.1.1 Domain Translation (DT) . . . 113
5.1.2 Spatial-Fourier Mutual Learning . . . 115
5.1.3 Hierarchical cross-encoder feature swapping (HCFS): . . . 116
5.1.4 Prompting-based Spatial-Fourier Interaction . . . 116
5.1.5 Loss Function . . . 117
5.2 Experiments . . . 118
5.2.1 Dataset Details . . . 118
5.2.2 Implementation Details . . . 118
5.2.3 Quantitative Results . . . 119
5.2.4 Qualitative Results . . . 120
5.2.5 Ablation Study . . . 121
5.2.6 Analysis of Parameters and FLOPs . . . 123
5.2.7 Scalability Verification . . . 123
5.2.8 Convergence analysis . . . 124
5.2.9 Application Test . . . 126
5.3 Summary . . . 126

6 Generating and Enhancing River-water Images **129**
6.1 Methodology . . . 131
6.1.1 Data Generation . . . 132
6.1.2 Proposed Image Enhancement Model . . . 133
6.1.2.1 Transformer Block . . . 134
6.1.2.2 Depth-degraded feature interaction Block . . . 134
6.2 Experiments . . . 135
6.2.1 Dataset Details . . . 135
6.2.2 Implimentation Details . . . 135
6.2.3 Quantitative Result . . . 136
6.2.4 Qualitative Result . . . 136
6.2.5 Ablation Study . . . 137
6.3 Summary . . . 138

7 Dual-domain Attention Network for Underwater Image Super-resolution 141
7.1 Attention-based Spatial-Frequency Information Network for Underwater Image Super-resolution . . . 142
7.1.1 Proposed Model . . . 143
7.1.1.1 Loss Function . . . 146
7.1.2 Experiments . . . 147
7.2 Efficient-USR: Prompt Guided Dual-domain Feature Information for Efficient Underwater Image Super-resolution . . . 150
7.2.1 Proposed Model . . . 152
7.2.1.1 Spatial-Frequency Interaction Block (SFIB) . . . 153
7.2.1.2 Prompt-guided Cross Attention Block (PCAB) . . . 155
7.2.2 Experiments . . . 156
7.2.2.1 Quantitative Results . . . 156
7.2.2.2 Qualitative Results . . . 157
7.2.2.3 Ablation Study . . . 157
7.3 Summary . . . 158

8 Conclusion and Future Work **161**
8.1 Conclusion . . . 161
8.2 Future Work . . . 163

Publications **165**

References **169**

List of Figures

1.1 Underwater imaging model. 2
1.2 Challenges of underwater imaging. 3
1.3 Graphical demonstration of underwater image enhancement problem. 4
1.4 Graphical demonstration of underwater image super-resolution problem. . . . 5
1.5 Impact of image enhancement on downstream computer vision tasks. 5

2.1 A graphical demonstration of multi-level wavelet sub-bands. 20
2.2 A graphical demonstration of degraded images and Fourier spectra of degraded images. 22
2.3 A graphical demonstration of physical model-based UIE. 23
2.4 A graphical demonstration of physical model-free UIE. 24
2.5 A graphical demonstration of learning-based UIE. 24
2.6 A graphical demonstration of interpolation-based methods. 26
2.7 A graphical demonstration of learning-based methods. 27
2.8 A graphical demonstration of the low contrast and poor visibility in the enhanced results of existing works. 28
2.9 A graphical demonstration of wavelength vs. receptive field size. 29
2.10 A graphical demonstration of single scale and multi-scale analysis. 29
2.11 A graphical demonstration of RGB and HSV color space. 30
2.12 A graphical demonstration of multiple degraded images. 31
2.13 A graphical demonstration of ocean and river images. 31
2.14 A graphical demonstration of the limitations of existing UISR methods. . . . 32

3.1 Visual comparisons of enhancement example on low contrast underwater image. The proposed Lit-Net model enhanced the contrast and removed the color deviation. 41
3.2 Overview of the proposed model for UIE. The model accepts a degraded underwater image as input and produces an improved image that is visually and spatially enhanced. 42

3.3 Network architecture of CBAM block [1]. 44
3.4 Network architecture of proposed encoder and decoder layer. 46
3.5 Visual demonstration against the state-of-the-art UIE models on UIEB dataset. 55
3.6 Visual demonstration against the state-of-the-art UIE models on SUIM-E dataset. 55
3.7 Visual demonstration against the state-of-the-art UIE models on EUVP dataset. 56
3.8 Visual demonstration against the state-of-the-art UIE models on UIEB challenge test set. 56
3.9 Visual demonstration on UIDEF dataset. 57
3.10 Visual demonstration of different modules of Lit-Net. (a) Degraded image, (b) MRAN without attention module, (c) MSAN without spatial attention module, (d) Lit-Net without attention module, (e) Lit-Net with fixed size kernel, (f) Lit-Net without channel-wise split, and (g) Full Lit-Net 57
3.11 Examples on foggy and low-contrast images. 58
3.12 Visual comparison of the semantic segmentation maps derived from the improved underwater images from four existing UIE works. 59
3.13 Visual demonstration of object detection by utilizing the improved images from three existing UIE works. 60
3.14 Visual demonstration of object detection by utilizing the improved images from three existing UIE works. 60
3.15 Architecture of X-CAUNET. The model accepts a degraded underwater image as input and produces a corresponding visually-enhanced one. R-G-B denote the three image channels, with color-coded arrows indicating the channel-specific paths. P = R or G channel. 62
3.16 Visual comparison with existing methods, with SSIM/PSNR values attained shown inside the red boxes. 66
3.17 (a) Color-channel dependencies in raw underwater images; (b) Impact on semantic segmentation by using the enhanced image produced by X-CAUNET as compared to previous UIE methods. 67
3.18 Ablation studies with different Loss Settings (LS) and different Component Settings (CS) of the proposed model. 67
3.19 Visual results of the proposed model with different component settings from the ablation study along with respective UIQM values obtained. 68

4.1 R-G-B channel-wise image visualization and cumulative distribution of degraded and enhanced (by proposed method) image. 72

4.2 Overview of the proposed UEnhancer, which utilizes the dual color space (RGB and HSV) and learns enriched local-global contextual feature representation for UIE. . . . 74
4.3 The components of the transformer block. MHA, SAE, and FFN indicate the multi-head attention, spatial attention enhancer, and feed-forward network, respectively. . . . 75
4.4 Illustration of proposed (a) Dual-color interaction (DCI) and (b) Latent space fusion with convolution attention (LSFCA) block. . . . 78
4.5 Qualitative comparison with SOTAs on EUVP dataset. . . . 86
4.6 Qualitative comparison with SOTAs on UIEB dataset. . . . 87
4.7 Qualitative comparison with SOTAs on SUIM-E dataset. . . . 87
4.8 Qualitative comparison with SOTAs on LSUI. . . . 87
4.9 Qualitative comparison with SOTAs on C60 and U45 datasets. . . . 87
4.10 Comparison of pixel distribution statistics of different methods through a pie chart. . . . 88
4.11 Intermediate feature maps, showing channel-specific focus on fine edges (red), textures (green), and haze patterns (blue) that are progressively fused in the decoder for accurate enhancement. . . . 88
4.12 Visual comparisons between proposed model and the existing model. The middle and right columns show the results of the existing and proposed methods, respectively. . . . 89
4.13 Visual effect of two losses. The top corner indicates the LPIPS score. . . . 89
4.14 Influence of various components in UEnhancer. The top corner indicates the LPIPS score. . . . 90
4.15 The display of keypoint matching. Note that the proposed method achieves the highest number of keypoint matches. . . . 90
4.16 Visual evaluation of semantic segmentation on SUIM dataset. . . . 91
4.17 Canny edge detection comparison between existing methods and proposed work. . . . 91
4.18 Top row represents the input and corresponding structure of the input. The last row represents the output and corresponding structure of output by the proposed model. . . . 92
4.19 Comparison of different scanning strategies for Mamba Block. Unlike (a) bidirectional, (b) raster, or (c) diagonal, proposed (d) A^* guided GIFH traversal adaptively follows degradation-aware paths, enabling more effective and sparse feature propagation. . . . 94

4.20 Overall Architecture of D2Mamba. The network extracts spatial and frequency features via a dual-stream encoder, adaptively traverses them using A* guided GIFH, enhances sequential features with Mamba blocks, and reconstructs images through a gated decoder. 95
4.21 Qualitative results on C60 and U45. D2Mamba method produces clearer structures and more natural colors compared to other existing methods. . . . 103
4.22 Enhanced image under various loss configuration. 106
4.23 Visualization of learned paths using GIFH. 107

5.1 Degraded images, Fourier spectra of degraded images, and enhanced images by PLUM. 110
5.2 (a) The heatmap shows the pixel-wise absolute error between the input image and the ground truth, as well as the enhanced image and the ground truth. (b) Trade-off between SSIM vs FLOPs (G) on UIEB dataset. 110
5.3 Overview of the proposed model with all the sub-components. 112
5.4 The first and second row show the enhancement results of existing and the proposed methods on SUIM-E dataset, while the third and fourth row illustrate the enhancement results on UIEB dataset. 121
5.5 Enhanced results on U45, C60, and UCCS dataset 122
5.6 Visual results of the ablation study. 124
5.7 Memory Usage of the proposed model from small to extra-large. 125
5.8 Convergence of the loss curve. 126
5.9 The first, third, fifth, and seventh row shows the enhancement results of existing and the proposed methods. The second row displays the edge detection result of the first row. The fourth row presents the depth map of the third row. The sixth row illustrates the segmentation result of the fifth row. The eighth row features the pie chart of colors of the seventh row. 128

6.1 Examples of degraded underwater images in the ocean (first row), the generated muddy images for the corresponding ocean images (second row), and the available ground truth images (third row). 130
6.2 Visual comparison of generated muddy water images with respect to different α values for a few sample images. 131
6.3 Overview of the proposed model along with all the sub-components. 131
6.4 Visual comparison of the proposed and SOTA methods. 137
6.5 Validation on real-world muddy water images. 137
6.6 Visual comparison with different settings in proposed model. 138

7.1 Visual demonstration of 3X super-resolve example on UFO-120 dataset. . . . 142

7.2 Overview of the proposed model for UISR. It accepts a degraded LR image as input and produces an SR image that is visually and spatially enhanced. s.H/s.W denotes the image height/width scaled by factor s. CA indicates the channel attention. 143
7.3 Network architecture of CBAM [1]. 145
7.4 Visual comparison of the proposed model with existing works on UISR. SSIM/UIQM score is marked on the corner. 147
7.5 Scale-wise effect of loss functions on UFO-120 dataset. 148
7.6 Visual demonstration of different components of the proposed model on UFO-120 (3x). PSNR is marked on the corner. 148
7.7 Visual comparison of L1 and color channel-specific L1 loss on UFO-120 dataset. PSNR/UIQM is marked on the corner. 149
7.8 Trade-off between PSNR vs GFLOPs on UFO-120 $\times 4$ dataset. The results show the superiority of EFFICIENT-USR among existing methods. 151
7.9 Overview of the proposed Efficient-USR along with all the sub-components. . 152
7.10 Visual comparison with SOTA methods on UFO-120 ($\times 2$) and USR-248 ($\times 2$) datasets. 154
7.11 PSNR value in each epoch on UFO-120 ($\times 2$). 157
7.12 Visual comparison of different settings of Efficient-USR. 158

List of Tables

3.1 MSE, PSNR, SSIM, MS-SSIM, LPIPS, UIQM, UISM, and BRISQUE comparison on EUVP dataset with the best-published works for UIE. First, second, and third best performance are represented in red, blue, and green colors, respectively. ↓ indicates lower is better. 49

3.2 PSNR, SSIM, MS-SSIM, LPIPS, UIQM, UISM, and BRISQUE comparison on UIEB test set with the best-published works for UIE. First, second, and third best performances are represented in red, blue, and green colors, respectively. ↓ indicates lower is better. 51

3.3 PSNR, SSIM, MS-SSIM, LPIPS, UIQM, UISM, and BRISQUE comparison on SUIM-E test set with the best-published works for UIE. First, second, and third best performances are represented in red, blue, and green colors, respectively. ↓ indicates lower is better 52

3.4 UISM, UIQM, and BRISQUE comparison on UIEB challenge set with the best-published works. First, second, and third best performances are represented in red, blue, and green colors, respectively. 52

3.5 UIQM comparison on UCCS and U45 datasets with the best-published works. First, second, and third best performances are represented in red, blue, and green colors, respectively. 53

3.6 Comparison on UIDEF dataset 53

3.7 Comparison on LSUI dataset 54

3.8 Comparison with the model parameters and GFLOPs of the SOTA model with the image size 256×256. Lower is better. 54

3.9 Effectiveness of various cost functions on three datasets: EUVP, UIEB, and SUIM-E, for the task of UIE 54

3.10 Effectiveness of various modules on two datasets: UIEB and SUIM-E for the task of UIE. 58

3.11 Ablation on UIEB dataset by interchanging the kernel sizes. 58

3.12 Mean average precision (mAP) comparison on detection task 59

3.13 UIQM comparison on U45 dataset. The first, second, and third best performances are represented in red, blue, and green, respectively. 65
3.14 Comparison with the state-of-the-art on three datasets across six different evaluation metrics. 65

4.1 Quantitative evaluation on E515 testset with the SOTA methods. Best, second best, and third best results are highlighted in red, blue, and green colors, respectively. 81
4.2 Quantitative evaluation on U90 test set with the SOTA works for UIE. . . . 83
4.3 UIQM comparison on C60, U45, and UCCS dataset with the SOTA works. . 84
4.4 Quantitative evaluation on S110 test set with the SOTA works for UIE. . . . 84
4.5 Quantitative evaluation on L400 test set with the SOTA works for UIE. . . . 86
4.6 Parameter comparison with SOTAs. 86
4.7 Effectiveness of various cost functions on U90. 88
4.8 Effectiveness of various settings in the proposed UIE model on U90. 89
4.9 PSNR and SSIM comparison on LOL dataset with existing methods. 92
4.10 Quantitative comparison of D2Mamba with existing methods on two full-reference datasets (UIEB, LSUI) and four no-reference datasets (U45, UCCS, C60, SQUID). The top-performing method is highlighted in red, the second-best in blue, and the third-best in green. (↑ higher is better, ↓ lower is better). 101
4.11 Comparison of quantitative results on D2Mamba obtained using different loss settings on the UIEB dataset. 105
4.12 Comparison of quantitative results on D2Mamba using various scanning types on the UIEB dataset. 105
4.13 Comparison of quantitative results on D2Mamba by varying path numbers and Mamba layers on the UIEB dataset. 105

5.1 SSIM, PSNR, and UIQM result of the proposed method and SOTA methods on the UIEB dataset. Red, blue, and green indicates the 1^{st}, 2^{nd}, and 3^{rd} best results. 119
5.2 SSIM and PSNR result of the proposed method and SOTA methods on the LSUI dataset. 119
5.3 SSIM and PSNR result of the proposed method and SOTA methods on the SUIM-E dataset. 120
5.4 UIQM results of the proposed method and SOTA methods on the U45, C60, and UCCS datasets. 120
5.5 Quantitative result of different components of the proposed model on the UIEB dataset. 123

5.6 Quantitative result of cross-dataset for UIE task. 123
5.7 Parameters and GFLOPs comparison of the proposed method and SOTA methods. 124
5.8 Scalability Verification on the UIEB dataset. 125

6.1 Quantitative result of proposed model and SOTA models. 136
6.2 Quantitative comparison of different settings of the proposed model. 138

7.1 Comparison between proposed model with the State-of-the-Art methods on UFO-120 dataset. Red, blue, and green indicate best, second-best, and third-best values, respectively. 146
7.2 Comparison between proposed model with the State-of-the-Art methods on USR-248 dataset. 150
7.3 Effect of different components of the proposed model on UFO-120 dataset with scaling factor 2. Red indicates best value 150
7.4 Quantitative comparison of the proposed approach with the state-of-the-art approaches on UFO-120 and USR-248 datasets. Red and blue indicate the best and second best results. 156
7.5 Effect of different components of the proposed Efficient-USR. 158

List of Algorithms

1 *MFS* loss's algorithm 81
2 Muddy Water Image Generation 133

List of Symbols

$*$ Convolution operation.

I_D Degraded image.

I_E Enhanced image.

I_{HR} Higher resolution image.

I_{LR} Low resolution image.

I_{SR} Super resolution image.

λ Weight factor.

$\mathbb{E}$ Expectation.

$\mathbb{R}$ Real number.

$\mathcal{L}$ Loss function.

$\odot$ Concatenation operation.

$\otimes$ Matrix multiplication.

$\prod$ Product operation.

$\sum$ Summation operation.

$c^{k\times k}$ Convolution with kernel size $k \times k$.

$||.||$ Norm of a vector.

List of Acronyms

AUV Autonomous Underwater Vehicle.

BRISQUE Blind/Reference-less Image Spatial Quality Evaluator.

CBAM Convolutional Block Attention Module.

CNN Convolutional Neural Network.

DFT Discrete Fourier Transform.

DL Deep Learning.

DNN Deep Neural Network.

DoG Difference of Gaussian.

DT domain translation.

DWT Discrete Wavelet Transform.

FFT Fast Fourier Transform.

GAN Generative Adversarial Network.

GDFFN Gated-Dconv Feedforward Network.

GIFH Geodesic Information-Field Heuristic.

HR Higher-resolution.

LoG Laplacian of Gaussian.

LPIPS Learned Perceptual Image Patch Similarity.

LR Low-resolution.

MLP Multi-Layer Perceptron.

MRAN Multi-resolution Attention Network.

MS-SSIM Multi-Scale Structural Similarity Index Measure.

MSAN Multi-scale Attention Network.

MSE Mean Squared Error.

NSS Natural Scene Statistics.

PCAB Prompt-guided Cross-attention Block.

PSNR Peak Signal-to-Noise Ratio.

ROV Remotely Operated Vehicle.

SAE Spatial Attention Enhancer.

SKFF Selective Kernel Feature Fusion.

SOTA state-of-the-art.

SR Super-resolution.

SSIM Structural Similarity Index Measure.

SSM State Space Model.

SWAL Spectral Wasserstein Attenuation Loss.

UIE Underwater Image Enhancement.

UIQM Underwater Image Quality Measure.

UIR Underwater Image Restoration.

UISM Underwater Image Sharpness Measure.

UISR Underwater Image Super-resolution.

"Exploration is really the essence of the human spirit."

~Frank Borman

1

Introduction

Images serve as visual representations of real-world scenes, with camera systems emulating aspects of human vision across various fields, including underwater imaging, marine observation, robotic sensing, and environmental research. In recent times, AUV based underwater imagery has become prevalent due to diverse industry and research applications such as monitoring the marine ecosystem, exploring energy resources, investigating the seabed, modern defense systems, and many more [2–4]. Figure 1.1 depicts the underwater imaging model. Underwater images suffer from various forms of visual degradation, with different regions of the image being affected at varying levels [5]. This degradation is caused by a variety of physical phenomena, such as suspended particles, plankton, and selective attenuation based on the light wavelength in the underwater environment, leading to a global nonuniform

shift in the feature representation [6]. As a consequence, underwater images exhibit low contrast and low visibility (refer to Figure 1.2), making it challenging to extract meaningful information from them. The absorption of the red component is significant, whereas the blue-green component encounters minimal absorption. The abundance of blue and green color channels in underwater imagery leads to the occurrence of the color casting problem. Due to this, underwater images typically have a dominant bluish and greenish color tone. The presence of these obstacles presents significant challenges for vision-based underwater operations [7]. For e.g., in underwater scenarios, it may be profoundly challenging for an AUVs to figure out what lies ahead and lead to a crashed navigation system. Consequently, it is crucial to employ advanced techniques to enhance the quality of underwater images and extract valuable information from these visually impaired conditions [8, 9]. In recent times, there has been a significant emphasis on enhancing underwater image quality to improve the visual appearance of underwater images [10].

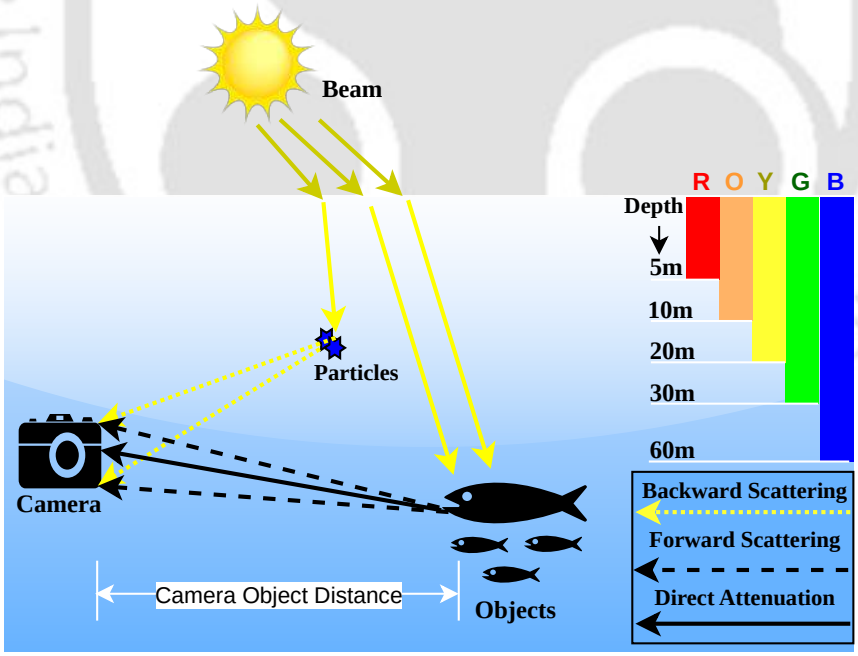


Figure 1.1: *Underwater imaging model.*

Although enhancement methods primarily address wavelength-dependent attenuation and scattering by improving color balance and contrast, they cannot recover the high-frequency spatial details that are lost during image formation. This shortcoming can hinder

tasks such as object classification [11], detection [12], and segmentation [13], which need clear texture and object boundaries. SR [14–17] helps solve this problem by adding high-frequency details and sharpening edges. This increases the spatial resolution and improves the overall quality of underwater images, especially for small or distant objects.

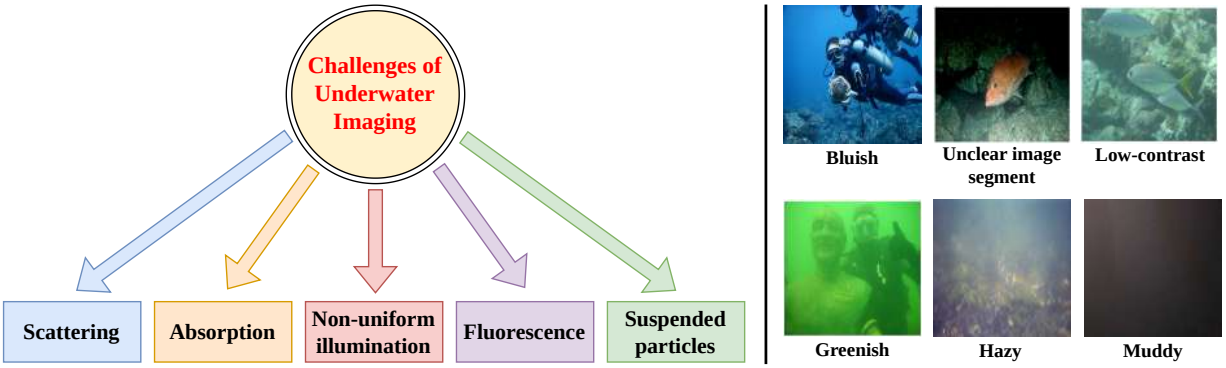


Figure 1.2: *Challenges of underwater imaging.*

In recent years, researchers have increasingly focused on DL-based approaches [18–24] due to their significant performance advantages over traditional, handcrafted techniques. Low-level vision tasks such as image enhancement and super-resolution are inherently ill-posed, as multiple high-quality solutions may correspond to a single degraded observation. Earlier methods attempted to address this issue by introducing handcrafted priors to transform the ill-posed problem into a more tractable one. However, such prior-based approaches often suffer from limited efficiency and poor generalization, as they fail to fully exploit the rich and diverse information present in real-world data, frequently leading to visual artifacts such as blocking and over-smoothing. In contrast, learning-based methods leverage large-scale data to implicitly learn complex image priors, enabling more robust restoration of textures and structures. The existing literature clearly demonstrates that DL-based models for image enhancement and SR achieve superior visual quality and generalization performance, while recent advances in network design and model optimization have also made these approaches suitable for deployment in resource-constrained and real-time environments.

Motivated by these observations, this thesis presents learning-based models for enhancing the quality and resolution of underwater images. In particular, this thesis focuses on two key

problems: (1) underwater image enhancement and (2) underwater image super-resolution. The following sections provide an overview of these challenges and describe the motivation behind the proposed data-driven solutions.

1.1 Underwater Image Enhancement

Given a degraded underwater image $I_D \in \mathbb{R}^{H \times W \times 3}$, the objective of UIE is to recover a visually improved image $I_E \in \mathbb{R}^{H \times W \times 3}$. This is achieved by learning a mapping function f that transforms the degraded input into its enhanced counterpart:

$$I_E = f(I_D) \tag{1.1}$$

where, H and W denote the height and width of the image. Figure 1.3 shows the qualitative demonstration of this task.

Figure 1.3: *Graphical demonstration of underwater image enhancement problem.*

1.2 Underwater Image Super-Resolution

Given a low-resolution underwater image $I_{LR} \in \mathbb{R}^{H \times W \times 3}$, the goal of SR is to generate a Higher-resolution (HR) image $I_{SR} \in \mathbb{R}^{s.H \times s.W \times 3}$. Using a Deep Neural Network (DNN) C(.), the super-resolved output can be formulated as:

$$I_{SR} = C(I_{LR}, s) \tag{1.2}$$

where H and W denote the spatial dimensions of the input image, and s represents the upscaling factor. Figure 1.4 shows the qualitative demonstration of this task.

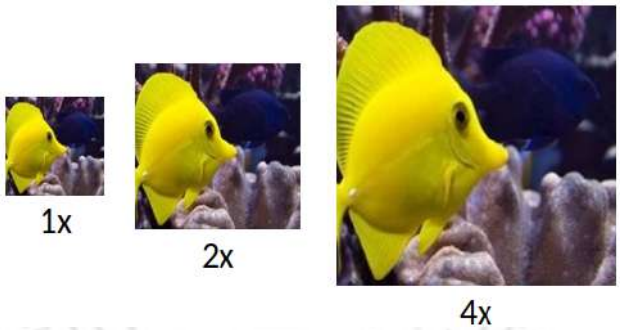


Figure 1.4: *Graphical demonstration of underwater image super-resolution problem.*

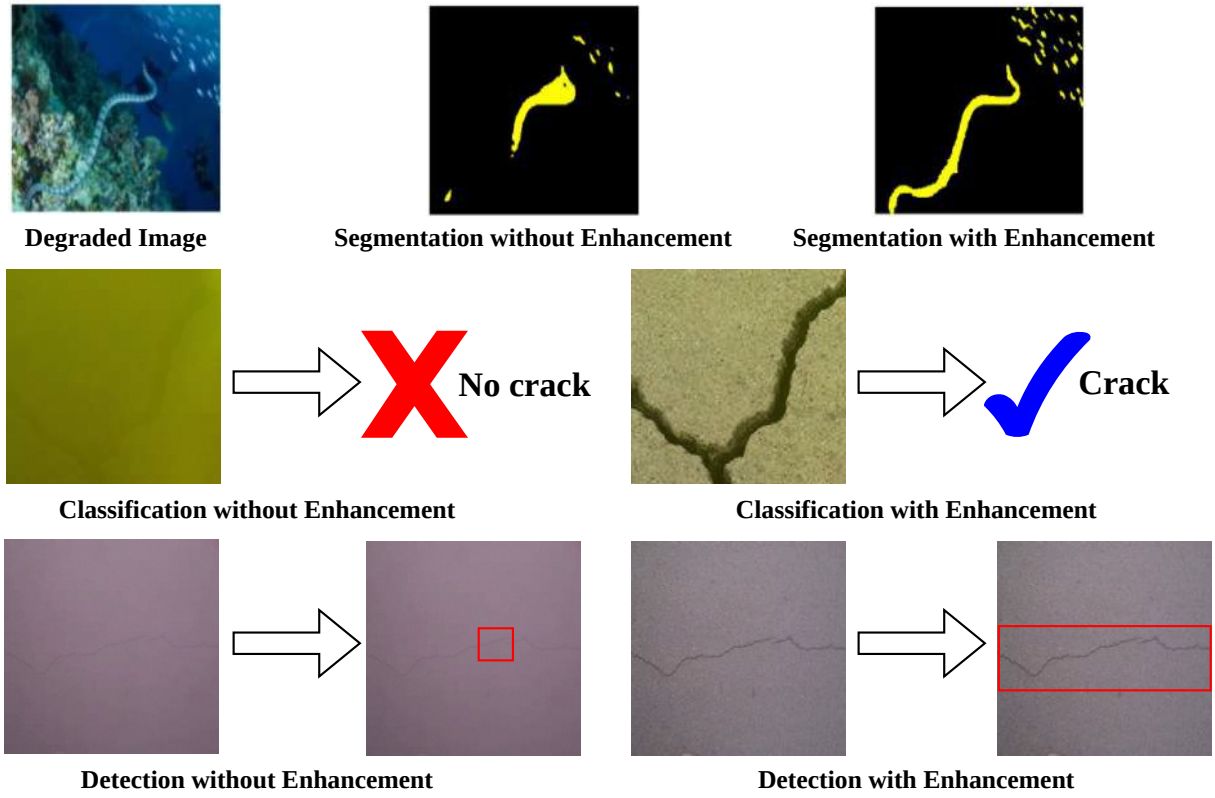


Figure 1.5: *Impact of image enhancement on downstream computer vision tasks.*

1.3 Applications

Improvements in low-level vision tasks, like enhancing images and increasing resolution, can significantly boost tasks such as object detection, classification, and semantic segmentation

(refer to Figure 1.5). This is especially important in underwater environments, where poor visibility can make autonomous systems less reliable. The underwater image enhancement and super-resolution models created in this thesis have several useful applications:

- **Underwater surveillance and security:** Harbor security, marine border surveillance, and underwater intrusion detection need dependable vision systems that work well in low visibility [25]. Clear and high-quality images help with spotting objects and identifying threats in murky water.
- **Archaeological and geological exploration:** Underwater archaeological sites, rock formations, and sediment deposits often look distorted because of light scattering and murkiness [26]. Enhancing the images helps to clearly document these structures.
- **Aquaculture and fisheries management:** Counting fish, monitoring their health, and analyzing their behavior in aquaculture needs clear underwater images. Better visuals help track fish populations, spot infections, and watch feeding activities in real time [27].
- **Search and rescue operations:** In underwater disaster situations, visibility is often very low, especially in muddy rivers or murky water. Clearer images help divers and robots find submerged objects and victims, even in difficult conditions [28].

1.4 Motivation of the Research Work

Over the years, notable advancements have been made in enhancing underwater images; however, current DL-based [29–33] methods still suffer from visual artifacts, including color distortion, low contrast, poor visibility, and haziness. These degradations have an adverse impact on the applications mentioned above. Existing methods overlook the fact that the visual distortions in underwater images vary depending on the wavelength of the light. Due to this, special consideration is required in each color channel. Several existing methods [33,34] utilize multiple color representations for UIE; most of them rely on shallow fusion strate-

gies that do not fully leverage the complementary properties of these color spaces. Also, frequency cues are not fully explored to guide spatial restoration, resulting in residual blur or over-smoothing. Most existing enhancement works [35–37] treat multiple degradations separately, thereby limiting their applicability in many computer vision tasks. As of now, existing works only consider the enhancement of oceanic scenarios, ignoring the riverine conditions where clay, silt, sand, and similar particles make these images muddy. In case of UISR, super-resolved images often lack rich texture details and high-frequency information, resulting in unrealistic or blurred reconstructions. The limitations of existing approaches for UIE and UISR motivated this research work.

1.5 Objectives of the Thesis

Motivated by the above observations, the main objectives of this dissertation are as follows:

- Develop deep learning-based models using color channel-aware feature learning for ocean-water image enhancement.
- Design deep models for ocean-water image enhancement using dual-color and dual-domain (spatial-frequency) feature transformations.
- Develop a multi-degradation feature learning-based deep model for ocean-water image enhancement.
- Devise a deep learning-based model for generating and enhancing river-water images.
- Design deep learning models for underwater image super-resolution utilizing spatial frequency domain features with limited computational resources.

1.6 Contributions of the Thesis

The major contributions of this dissertation are as follows:

1.6.1 Color Channel-Aware Feature Learning for Ocean-water Image Enhancement

In the first chapter, two models are proposed that use color channel-aware feature learning to improve the quality of underwater ocean images. These approaches leverage the unique properties of individual color channels to effectively tackle problems like color distortion, low contrast, and spatial inconsistencies often found in underwater environments. The first work presents a lightweight CNN-based multi-stage model called Lit-Net, which is designed to perform multi-resolution and multi-scale analysis for underwater image enhancement. A new encoder block is introduced, using parallel 1×1 convolution layers to capture fine local details while ensuring efficiency. Additionally, a modified weighted color channel-specific L1 loss (cL1) is incorporated to enhance the recovery of color accuracy and subtle image details. Building upon previous work, the second work proposes X-CAUNET, a transformer-based framework that leverages a message-passing mechanism to capture cross-channel correlations.

1.6.2 Dual Feature Transformations for Ocean-water Image Enhancement

Following the progress made via color channel-specific feature learning, the next step is to integrate richer representations that span multiple color spaces and both spatial and frequency domains. In the second contributory chapter, two main contributions have been made that achieve this through dual-color interaction and dual-domain feature modeling for more comprehensive ocean-water image enhancement.

First, UEnhancer, a Transformer-based model explicitly proposed by incorporating a dual-color (RGB-HSV) fusion-based hierarchical encoder-decoder architecture. It integrates a Spatial Attention Enhancer (SAE) within the transformer block to grasp the local context, which is crucial for the enhancement task. To preserve structural fidelity and high-frequency details, the training objective incorporates a multi-scale frequency-based SSIM loss.

Second, D2Mamba adopts a dual-domain information (spatial and frequency) with State Space Models (SSMs), enabling efficient global context modeling while preserving local details. Unlike conventional SSMs that rely on raster, bidirectional, cross, or diagonal scans, D2Mamba uses an A* search guided by physics-based Geodesic Information-Field Heuristic (GIFH) scan for feature traversal based on input degradation characteristics. GIFH combines feature gradients, high-frequency heterogeneity, and low-frequency semantic distance to compute adaptive costs, enabling the capture of both spatial and spectral dependencies. Further, a SWAL is introduced to enforce distributional alignment in the spectral domain, enabling perceptually consistent and physically consistent color restoration in enhanced underwater images.

1.6.3 Multi-Degradation Learning Framework for Oceanic Image Enhancement

While the previous approaches improved performance for typical types of degradation present in the dataset, real underwater environments often exhibit multiple degradations simultaneously, such as haze, low contrast, blur, and color distortion. Models designed for a single type of degradation often struggle with diverse conditions, which limits their effectiveness in real scenarios. To address this issue, the third contributory chapter presents a unified multi-degradation enhancement model that can handle various underwater degradations simultaneously. The model generates various degraded versions (haze, blur, low contrast, and color distortion) of the input degraded image through domain translation. Then, the original degraded and translated images are input into a multi-degraded enhancement network. Extensive experiments have demonstrated that this strategy significantly improves the model's performance on cross-dataset evaluation.

1.6.4 Generating and Enhancing River-water Images

Underwater imaging in rivers poses unique challenges due to clay, silt, sand, and other particles, which result in blurred and highly nonuniform scenes. To tackle these challenges, the

fourth contributory chapter presents the River-GEM model. River-GEM tackles two major issues found in river image enhancement: (1) the lack of datasets that accurately represent the muddy water conditions found in river environments and (2) the limited effectiveness of current enhancement models in dealing with the unique challenges of muddy water images. River-GEM includes (1) a dataset of highly degraded muddy water images generated using a principal component analysis-based fusion technique and (2) a novel enhancement model utilizing the depthmap guidance to learn complex contextual details of muddy water images. This dataset and enhancement network mark a significant step forward in this area of research.

1.6.5 Dual-Domain Attention Model for Underwater Image Super-Resolution

While UIE enhances color balance and contrast, it does not recover the fine spatial details lost during image acquisition. As a result, retrieving high-resolution structural information is vital for underwater vision systems. This thesis also explores UISR in the fifth contributory chapter, presenting two works that address the specific challenges of this task.

The first work proposes a DL-based UISR model that incorporates spatial information, as well as the transformed (wavelet) coefficients of degraded LR underwater images, through intelligent feature management. To ensure the visual quality of the super-resolved image, color channel-specific L1 loss, perceptual loss, and DoG loss are used in tandem with Structural Similarity Index Measure (SSIM) loss.

Recent advancements in deep learning have improved UISR, but complex architectures are often too computationally intensive for low-power devices, such as AUVs and Remotely Operated Vehicles (ROVs). To overcome this, Efficient-USR is proposed as a lightweight model that combines dual-domain information to capture both the global and local context by operating spatial-channel cross-attention on spatial-frequency features. Further, a Prompt-guided Cross-attention Block (PCAB) is introduced to efficiently encode degradation information in the prompt component and perform feature fusion across multiple

scales via cross-attention. Extensive experiments on two UISR benchmarks demonstrate the effectiveness of Efficient-USR.

1.7 Organization of the Thesis

The thesis is organized as follows:

- **Chapter 1: Introduction**

  Chapter one discusses underwater image enhancement and super-resolution. It describes why this research is important, outlines the goals of the thesis, and highlights its contributions.

- **Chapter 2: Research Background and Literature Survey**

  This chapter provides an overview of preliminary concepts and reviews previous work on the two main tasks: UIE and UISR.

- **Chapter 3: Color Channel-Aware Feature Learning for Ocean-water Image Enhancement**

  This chapter has two sections: a CNN-based model and a transformer-based model for enhancing ocean-water images by leveraging the unique properties of individual color channels.

- **Chapter 4: Dual Feature Transformations for Ocean-water Image Enhancement**

  This chapter first presents a dual-color transformer design that enhances ocean-water images by combining information from different color spaces. Next, it introduces a dual-domain state-space model that enhances the image in challenging ocean conditions.

- **Chapter 5: Multi-Degradation Feature Learning for Ocean-water Image Enhancement**

  This chapter describes a model that simultaneously addresses multiple underwater

degradations (e.g., haze, low contrast, blur, color distortion), enhancing its effectiveness in real ocean conditions.

- **Chapter 6: Generating and Enhancing River-water Images**

  This chapter presents a synthetic dataset for the riverine scenario and includes a model to improve the distorted river images.

- **Chapter 7: Dual-Domain Attention for Underwater Image Super-Resolution**

  This chapter presents a deep learning-based model for underwater image super-resolution that integrates DWT and FFT domain coefficients.

- **Chapter 8: Conclusion and Future Work**

  This chapter summarizes the main contributions of the thesis and suggests possible areas for future research.

"Any sufficiently advanced technology is indistinguishable from magic."

~Arthur C. Clarke

2

Research Background and Literature Survey

This chapter provides an overview of the foundational concepts relevant to this thesis. It first introduces the fundamental principles of deep CNNs, as well as Transformers and State-Space Models. Next, it introduces key image transformation techniques, including the 2D FFT, 2D DWT, and DoG. These serve as the foundation for the development of proposed deep learning–based models for UIE and SR. Then, it discusses the literature survey for UIE and UISR. Additionally, this chapter outlines the image quality evaluation metrics used to assess the proposed approaches, as well as the benchmark datasets employed for experimental validation.

2.1 Convolutional Neural Network

CNNs [38] are a class of DL models widely used in computer vision tasks due to their ability to effectively capture local spatial patterns and hierarchical features from images. CNNs employs convolutional filters to extract low-level features such as edges and textures in early layers, and higher-level semantic representations in deeper layers. Their parameter sharing and sparse connectivity make them computationally efficient and well-suited for image restoration tasks such as underwater image enhancement and super-resolution.

2.1.1 Standard Convolution

In standard convolution, a set of learnable kernels is convolved with the input feature map to produce output feature maps. For an input feature map $X \in \mathbb{R}^{H \times W \times C}$, the convolution operation can be expressed as:

$$Y_{i,j,k} = \sum_{c=1}^{C} \sum_{m,n} X_{i+m,j+n,c} W_{m,n,c,k}, \tag{2.1}$$

where W denotes the convolutional kernel. Standard convolution effectively captures local spatial context but incurs high computational cost when the number of channels increases.

2.1.2 Depthwise Separable Convolution

Depthwise separable convolution [39] decomposes standard convolution into two operations: depthwise convolution and pointwise convolution. The depthwise convolution applies a single filter per input channel, followed by a 1×1 pointwise convolution to combine channel information. This significantly reduces computational complexity and parameters while maintaining performance, making it suitable for lightweight and real-time models.

2.1.3 Dilated Convolution

Dilated (or atrous) convolution introduces a dilation rate to the convolutional kernel, allowing it to capture a larger receptive field without increasing the number of parameters. The dilated convolution operation is defined as:

$$Y(i,j) = \sum_{m,n} X(i + r \cdot m, j + r \cdot n)W(m,n), \tag{2.2}$$

where r is the dilation rate. This operation is effective for capturing multiscale contextual information, which is important for restoring large structures in degraded images.

2.1.4 Deformable Convolution

Deformable convolution [40] extends standard convolution by introducing learnable offsets to the sampling locations of the convolutional kernel. This allows the network to adaptively adjust its receptive field based on image content, enabling better modeling of geometric variations and object boundaries commonly found in underwater scenes.

2.1.5 Strip Convolution

Strip convolution [41] employs elongated kernels (e.g., $1 \times k$ or $k \times 1$) to efficiently capture long-range dependencies along horizontal or vertical directions. This operation is particularly effective for modeling structured patterns and directional features, while maintaining low computational complexity.

2.2 Transformer

Transformers [42] are attention-based models originally developed for sequence modeling and later adapted to vision tasks. Unlike CNNs, transformers rely on self-attention mechanisms to capture long-range dependencies and global context. Given a set of input tokens, the self-attention mechanism computes relationships between all token pairs using query, key,

and value projections:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{2.3}$$

where d is the feature dimension. In image processing, transformers are effective for global reasoning, color consistency modeling, and capturing long-range contextual relationships, making them well-suited for underwater image enhancement tasks.

2.3 State Space Model

SSMs [43] are sequence modeling frameworks that represent system dynamics using latent states evolving over time or space. An SSM can be defined by the following equations:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t, \tag{2.4}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t, \tag{2.5}$$

where $\mathbf{x}_t$ denotes the latent state, $\mathbf{u}_t$ is the input, and $\mathbf{y}_t$ is the output. Recent advances have adapted SSMs for vision tasks by modeling spatial dependencies efficiently with linear complexity. These models provide an effective alternative to transformers for capturing long-range dependencies with reduced computational overhead, making them attractive for high-resolution image restoration tasks.

2.4 Activation Functions

Activation functions introduce non-linearity into neural networks during feature map computation, allowing the network to model complex non-linear relationships. They also regulate the output range of neurons, which improves training stability, accelerates convergence, and mitigates issues such as vanishing or exploding gradients. Depending on their character-

istics, different activation functions offer trade-offs in sparsity, smoothness, and gradient propagation. Some widely used activation functions are described below:

- **Rectified Linear Unit (ReLU):**

$$f(x) = \max(0, x)$$

ReLU is computationally efficient and promotes sparse activations but may suffer from the dying ReLU problem for negative inputs.

- **Leaky Rectified Linear Unit (Leaky ReLU):**

$$f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{otherwise} \end{cases}$$

where α is a small constant (typically 0.001), allowing a small gradient for negative inputs.

- **Parametric Rectified Linear Unit (PReLU):**

$$f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{otherwise} \end{cases}$$

where α is a learnable parameter optimized jointly with the model parameters, enabling adaptive negative slope learning.

- **Sigmoid:**

$$f(x) = \frac{1}{1 + e^{-x}}$$

Sigmoid maps inputs to the range $(0, 1)$ and is commonly used in probabilistic modeling, though it is prone to vanishing gradients.

- **Hyperbolic Tangent (TanH):**

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

TanH maps inputs to $(-1, 1)$ and provides zero-centered activations, often leading to faster convergence than Sigmoid.

- **Gaussian Error Linear Unit (GELU):**

$$f(x) = x \cdot \Phi(x)$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution. GELU provides smooth probabilistic gating and is widely adopted in Transformer-based architectures.

- **Sigmoid Linear Unit (SiLU / Swish):**

$$f(x) = x \cdot \sigma(x)$$

where $\sigma(x)$ is the Sigmoid function. SiLU is smooth and non-monotonic, offering improved gradient flow and performance in deep networks.

2.5 Pooling Layer

Pooling is applied to two-dimensional feature maps to reduce their spatial resolution while preserving salient structural and contextual information. By aggregating local neighborhoods, pooling enhances translation invariance, reduces sensitivity to local noise, and decreases computational complexity. This operation is particularly useful in underwater imaging, where spatially varying distortions caused by scattering and turbidity are common.

- **Max Pooling:** Selects the maximum value within a pooling window:

$$f(x) = \max_{(i,j)\in\mathcal{R}} x_{i,j}$$

where $\mathcal{R}$ denotes a rectangular pooling region. This operation preserves strong activations such as edges and fine structures.

- **Average Pooling:** Computes the average value within a pooling window:

$$f(x) = \frac{1}{|\mathcal{R}|} \sum_{(i,j)\in\mathcal{R}} x_{i,j}$$

Average pooling retains global contextual information and smooths feature responses, making it more robust to noise.

- **Global Average Pooling (GAP):** Aggregates each feature map into a single scalar by averaging across all spatial locations:

$$f(x) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j}$$

where H and W denote the height and width of the feature map. GAP significantly reduces the number of parameters and helps prevent overfitting while maintaining global scene information.

2.6 2D Discrete Wavelet Transform

In image processing, the DWT decomposes an image into sub-band coefficients that correspond to different frequency components at multiple scales, allowing for separate analysis of coarse structures and fine details, as shown in Figure 2.1.

For a two-dimensional image $I(a, b)$, the DWT is applied independently along the horizontal and vertical directions. A single-level 2D DWT decomposition produces four sub-bands corresponding to different frequency components:

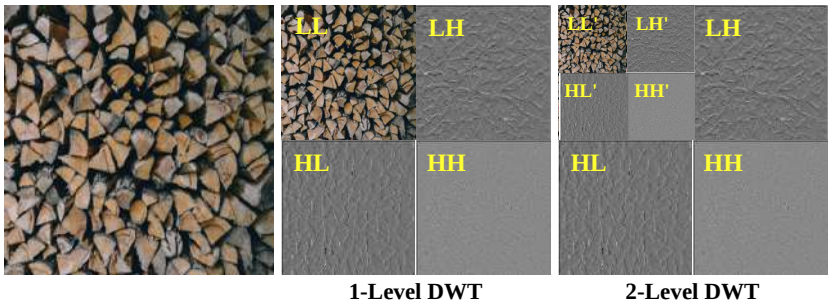


Figure 2.1: *A graphical demonstration of multi-level wavelet sub-bands.*

- **LL (Low-Low):** Low-frequency components in both directions, representing the coarse image structure.
- **LH (Low-High):** High-frequency components in the vertical direction, capturing horizontal edge details.
- **HL (High-Low):** High-frequency components in the horizontal direction, capturing vertical edge details.
- **HH (High-High):** High-frequency components in both directions, representing diagonal and fine texture details.

Mathematically, the 2D DWT decomposition can be expressed as successive filtering and downsampling operations along both dimensions:

$$LL = (I * L_a * L_b) \downarrow 2, \quad LH = (I * L_a * H_b) \downarrow 2, \tag{2.6}$$

$$HL = (I * H_a * L_b) \downarrow 2, \quad HH = (f * H_a * H_b) \downarrow 2, \tag{2.7}$$

where $*$ denotes the convolution operation, L and H represent the low-pass and high-pass wavelet filters, respectively, and $\downarrow 2$ indicates downsampling by a factor of two.

2.7 2D Fast Fourier Transform

Images are generally represented in the spatial domain, where pixel intensities define deviations in illumination across two spatial dimensions. Although spatial-domain representations are intuitive, numerous essential image features, such as texture, edges, and periodic patterns, are more effectively explored in the frequency domain. The 2D FFT transforms an image from the spatial domain into its frequency-domain representation by decomposing it into a weighted sum of sinusoidal basis functions with different spatial frequencies and orientations. This transformation shows how image intensity changes across different frequency components.

Mathematically, for a 2D image $I(a,b)$ of size $R \times C$, the 2D Discrete Fourier Transform (DFT) is defined as:

$$F(x,y) = \sum_{a=0}^{R-1} \sum_{b=0}^{C-1} I(a,b)\, e^{-j2\pi\left(\frac{xa}{R}+\frac{yb}{C}\right)} \tag{2.8}$$

where, (a,b) are spatial coordinates, (x,y) are frequency coordinates, and $j = \sqrt{-1}$.

The inverse 2D FFT reconstructs the spatial-domain image from its frequency-domain representation by combining both magnitude and phase information.

$$I(a,b) = \frac{1}{RC} \sum_{x=0}^{R-1} \sum_{y=0}^{C-1} F(x,y)\, e^{j2\pi\left(\frac{xa}{R}+\frac{yb}{C}\right)} \tag{2.9}$$

The FFT is an efficient algorithm for computing the DFT. While direct computation of the DFT requires $O(R^2C^2)$ operations, the FFT reduces the computational complexity to $O(RClog(RC))$.

An example of magnitude spectrums for a given input images has been shown in Figure 2.2.

2.8 Difference of Gaussian.

The DoG is a classical image processing technique widely used for edge detection and feature enhancement. It serves as an efficient approximation to the Laplacian of Gaussian (LoG)

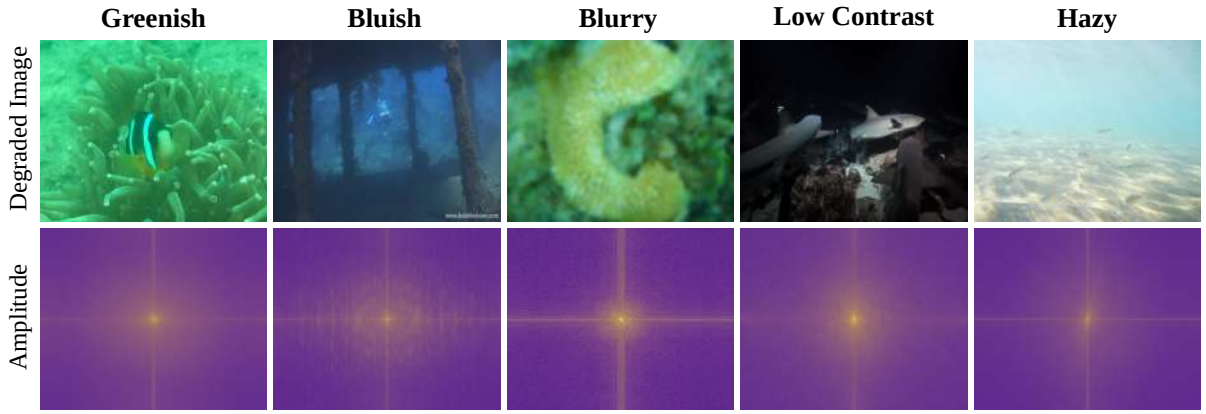


Figure 2.2: *A graphical demonstration of degraded images and Fourier spectra of degraded images.*

operator by subtracting two Gaussian-smoothed versions of an image computed at different scales. The DoG highlights regions of rapid intensity variation while suppressing noise and low-frequency background information.

Gaussian Filtering. A Gaussian filter smooths an image by reducing high-frequency noise and small-scale intensity variations. The two-dimensional Gaussian function is defined as:

$$G(a, b, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{a^2 + b^2}{2\sigma^2}\right), \tag{2.10}$$

where σ denotes the standard deviation controlling the scale of smoothing.

Given an input image $I(a, b)$, Gaussian smoothing is performed as:

$$L(a, b, \sigma) = I(a, b) * G(a, b, \sigma), \tag{2.11}$$

where $*$ represents the convolution operation.

Difference of Gaussian Operation. The DoG operator is obtained by subtracting two Gaussian-smoothed images computed using different standard deviations σ_1 and σ_2, with $\sigma_2 > \sigma_1$:

$$\text{DoG}(a, b) = L(a, b, \sigma_1) - L(a, b, \sigma_2). \tag{2.12}$$

This operation acts as a band-pass filter that enhances edges and fine structures while suppressing low-frequency illumination variations and high-frequency noise.

2.9 Literature Survey

In this section, an overview of prior research is provided in two main areas: (1) underwater image enhancement and (2) underwater image super-resolution.

2.9.1 Underwater Image Enhancement

In recent times, bottom-up and top-down methods have been suggested to enhance the visual qualities of underwater images. The bottom-up methods can be categorized into physical model-based and model-free. The evolution of top-down methods has introduced a pioneering approach for tackling low-level vision tasks, especially by leveraging deep learning frameworks to enhance the underwater image quality via data-driven strategies.

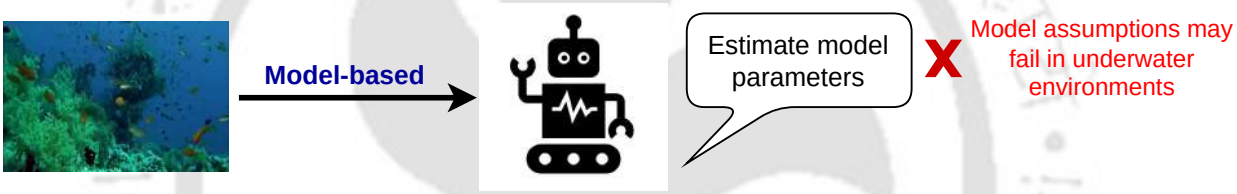


Figure 2.3: *A graphical demonstration of physical model-based UIE.*

2.9.1.1 Physical Model-based methods

Model-based methods [44–52] depend on the underwater imaging model, where specific priors are employed to approximate the background light and transmission map. The priors include underwater dark channel prior [47, 53], red channel prior [49, 54], haze-lines prior [45, 55], minimum information prior [48], blurriness prior [50], light attenuation prior [37, 56] etc. While these methods prove effective in enhancing underwater images under certain scenarios, the diverse array of underwater imaging degradations poses challenges in accurately estimating intricate model parameters. Consequently, the majority of these methods tend to exhibit sensitivity toward the specific types of underwater imaging degradation present. Thus, an inaccurate prior may lead to unstable and visually unpleasing artifacts in some regions of the enhanced image.

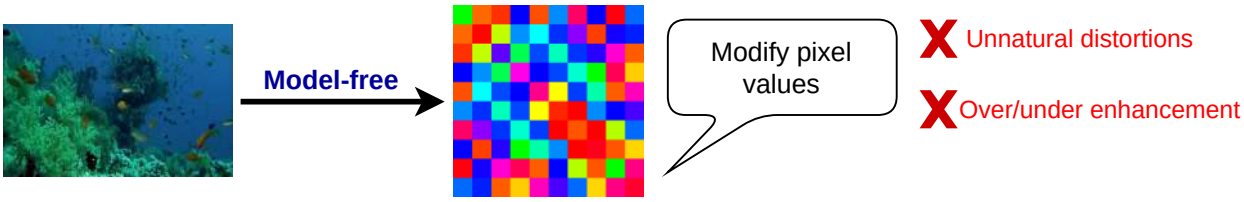


Figure 2.4: *A graphical demonstration of physical model-free UIE.*

2.9.1.2 Physical Model-free methods

Model-free methods are primarily focused on enriching underwater images' visual quality through direct manipulation of the pixel values. Commonly employed techniques consist of pixel distribution adjustment [57–62], retinex decomposition [9, 63–67], and fusion-based methods [57, 68, 69]. These model-free methods have the capability to improve the visual quality of degraded underwater images to some degree. However, by neglecting the underwater imaging process, model-free methods fail to deliver high-quality outcomes and introduce unnatural artifacts, leading to over- or under-enhancement in complex underwater scenarios.

Apart from the above methods, hardware-based methods like polarization filters [70] and special cameras [30, 71] can improve underwater images, but they need specific tools and are hard to use, especially since it's tough to take multiple shots of the same scene.

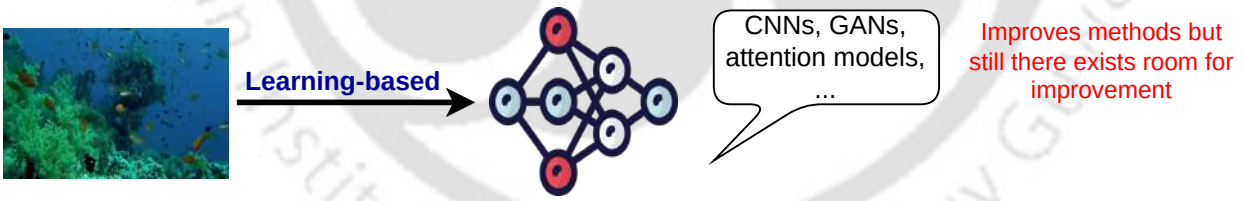


Figure 2.5: *A graphical demonstration of learning-based UIE.*

2.9.1.3 Deep Learning-based methods

In recent times, learning-based methods [18, 30, 72–80] have demonstrated significant potential in UIE, utilizing the capabilities of deep neural networks to model complex image degradations and recover visually and structurally improved images. These methods typically utilize CNNs, Generative Adversarial Networks (GANs), or transformer-based models to enhance image quality automatically.

The CNN-based algorithm enhances underwater images more effectively by matching raw images with their corresponding ground truth. The network blocks are built using various convolution operations, including naive convolution [18, 81], residual connections [72, 82], multi-stage [83] and multiscale fusion [84] for UIE.

GAN-based methods enhance underwater images by leveraging generator–discriminator interactions. One approach utilized GANs to synthesize paired data specifically for color correction [19]. A weakly supervised learning framework was proposed to address the scarcity of paired training samples [31, 85]. Improvements to the generator network have been made through multiscale cascading and residual learning strategies [86]. Another method introduced a dataset comprising both paired and unpaired images, trained using conditional GANs [32]. Additional techniques have incorporated underwater imaging models and joint learning mechanisms to refine color and structure restoration [87, 88]. GANs have also been applied with two-level depth estimation to enhance image quality [89].

The Transformer architecture is widely adopted in natural language processing and computer vision. UWAGA highlights the benefits of an automatic input channel grouping mechanism, leading to an adaptive group attention operation in the Swin Transformer [90]. [74] offers a multiscale feature fusion transformer block, while Spectroformer [77] presents a multi-domain query cascaded transformer that considers both localized and global information. However, these architectures often have higher computational costs than traditional convolution operations. Zhou et al. [91] proposed a physically guided spatial-frequency interaction Transformer for UIE, while [92] introduced UIEFormer, based on DehazeFormer and effective for small underwater image datasets, and Qing et al. [93] developed a transformer model leveraging HSV (Hue, Saturation, Value) and gamma correction for UIE.

SSMs [94], initially developed for control theory, have been adapted for deep learning due to their effectiveness in modeling sequential dependencies [95, 96]. Their linear complexity makes them a strong alternative to recurrent and attention-based architectures. Mamba [97] enhances SSMs with a selective mechanism and has been applied in computer vision for tasks like remote sensing classification [98] and medical image segmentation [99], as well

as UIE [100]. Lin et al. [101] introduced PixMamba, combining EMNet for global context and PixNet for detail enhancement. Guan et al. [102], Dong et al. [103], and Peng et al. [104] developed WaterMamba, O-mamba, and SS-UIE, respectively, to tackle challenges in underwater imaging and improve multi-scale information interaction.

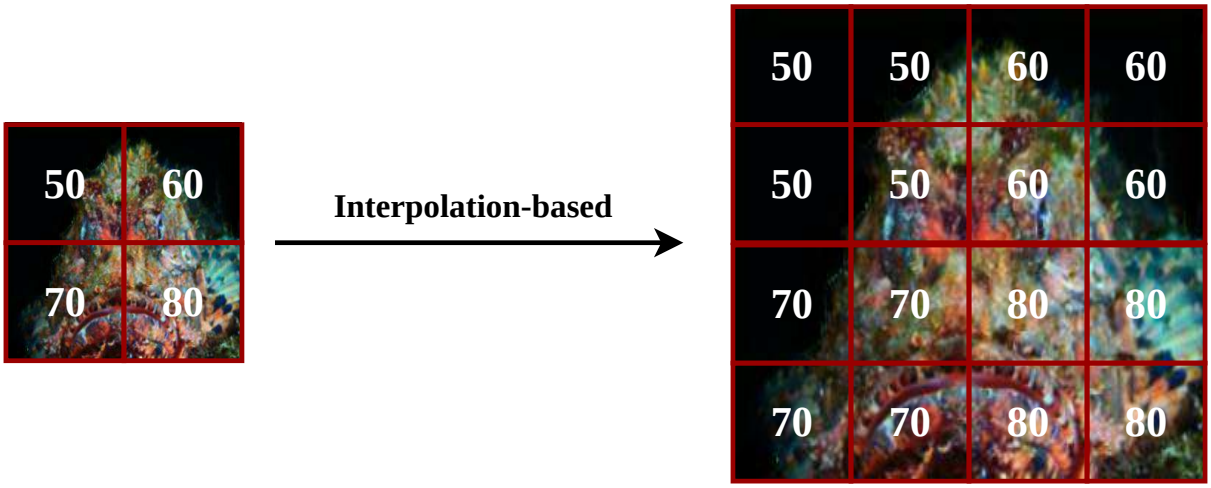


Figure 2.6: *A graphical demonstration of interpolation-based methods.*

2.9.2 Underwater Image Super-Resolution

The goal of UISR is to enhance the resolution of a low-quality image by generating a higher resolution version of that image [14–17]. The literature of UISR can be divided into two categories: (1) interpolation-based methods and (2) deep learning-based methods.

2.9.2.1 Interpolation-based methods

Interpolation-based methods, including bicubic, bilinear, and nearest-neighbor, use the values of adjacent pixels to infer the values of missing or unknown pixels [105]. These methods can introduce blurring and artifacts, particularly in areas with rapid changes in intensity or sharp edges, due to their reliance on averaging or smoothing neighboring pixel values.

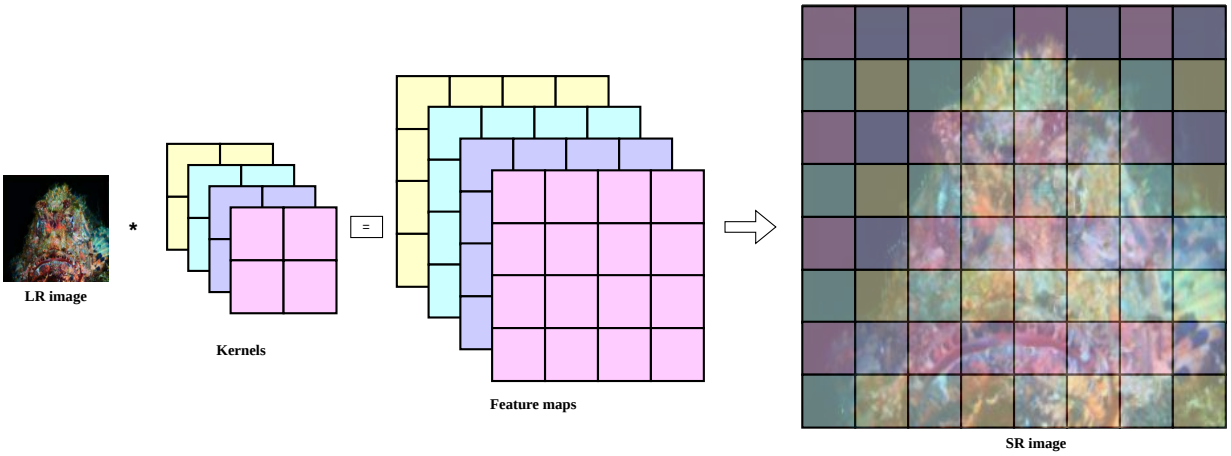


Figure 2.7: *A graphical demonstration of learning-based methods.*

2.9.2.2 Deep Learning-based methods

In current research, learning-based techniques have achieved SOTA performance by harnessing their powerful representational capabilities. Li et al. [30] proposed SRCNN, the initial SR model utilized a Color Features-based methodology and Multi-Layer Perceptron (MLP) technique. The work in [106] proposed an SR network based on wavelet transform that decomposes an image into a series of wavelets that vary in scale and frequency and improve dense block structure. Lu et al. [107] initially offered a self-similarity SR algorithm to the LR images for HR images. Finally, they got the SR images by applying a convex fusion rule that combines multiple estimates or observations into a single estimate of the HR images. The authors of [108] build a multi-modal optimization technique for underwater super-resolution that quantifies the perceptual quality of the image based on factors such as global content, color, and local style information to create an adversarial training pipeline. Islam et al. [109] presented a generative framework based on residual-in-residual networks that simultaneously improve and super-resolving underwater imagery for better visual perception. Sharma et al. [72] presented a deep learning-based framework using CNN, which consists of multiple stages, and it employs the distinct size of the kernel to each R-G-B color channel of an image, guided by its wavelength for concurrent UIE and UISR task. Recently, Zang et al. [110] proposed a multi-path network guided by an attention module that mainly utilizes the cross-convolution technique for UISR. Several methods have emerged

utilizing Transformers [111, 112], each effectively restoring intricate details and textures in images through their advanced feature representation capabilities. Despite significant improvements in UISR methods, they still suffer from visual artifacts in the super-resolved images, such as color distortion, lack of texture representation, and loss of detail.

2.10 Limitations of existing works

Over the past few years, numerous methods have been proposed for UIE and UISR, motivated by advances in learning-based modeling. Each approach introduces a distinct strategy to address underwater degradation by considering factors such as wavelength-dependent attenuation, scattering effects, and the statistical characteristics of underwater imagery. Despite these contributions, the existing literature on UIE and UISR still exhibits several limitations, which are discussed below.

1. With respect to underwater image enhancement, the current literature has the following limitations:
    - Despite notable progress, there is room for a lot of improvement in mitigating visual artifacts such as low contrast, poor visibility, color degradation, and others (see Figure 2.8).

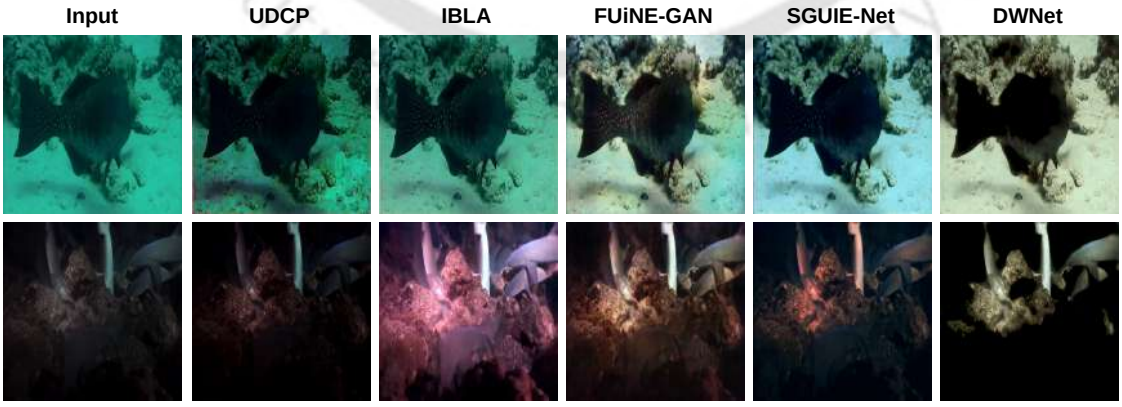


Figure 2.8: *A graphical demonstration of the low contrast and poor visibility in the enhanced results of existing works.*

- The majority of the existing works used an in-air enhancement model for UIE. Instead of the direct plug-and-play of in-air models, a dedicated model for UIE is required that considers the challenges present in the underwater environment.
- Underwater color degradation is highly wavelength-dependent, causing unequal distortion across R, G, and B channels. Treating all channels uniformly fails to recover true color relationships present in natural underwater illumination (see Figure 2.9).

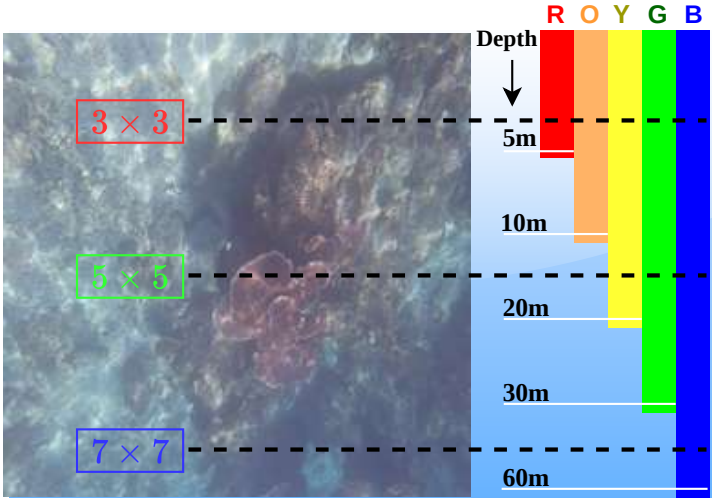


Figure 2.9: *A graphical demonstration of wavelength vs. receptive field size.*

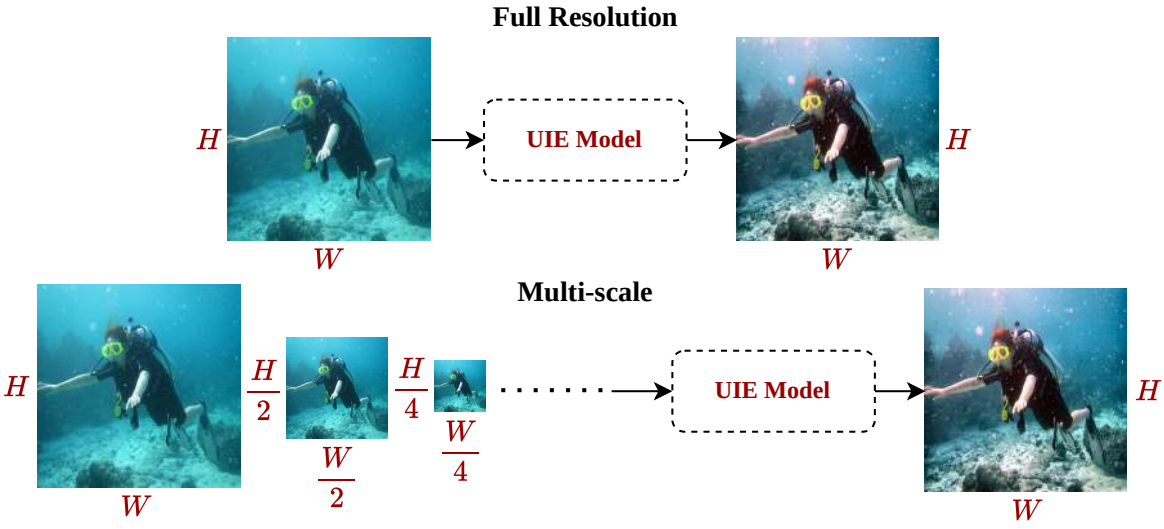


Figure 2.10: *A graphical demonstration of single scale and multi-scale analysis.*

- Earlier methods either analyze the input image at full resolution, resulting in spatial richness but contextual weakness, or progressively from high to low resolution, yielding reliable semantic information but reduced spatial accuracy (see Figure 2.10).
- Most previous methods utilized the RGB color space for the UIE task. Nevertheless, multiple color spaces allow the model to address color distortion, improve contrast and visibility, and preserve the natural appearance of underwater scenes (see Figure 2.11).

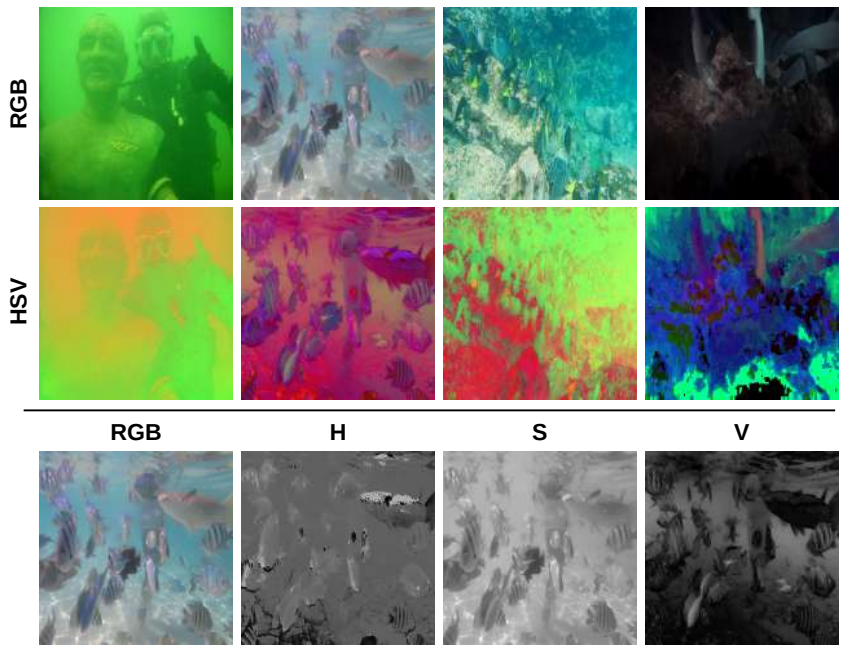


Figure 2.11: *A graphical demonstration of RGB and HSV color space.*

- Many methods rely solely on spatial-domain reasoning and overlook frequency-domain distortions. High-frequency attenuation, backscatter noise, and spectral imbalance remain unresolved under such approaches.
- Real underwater environments contain mixed artifacts such as haze, blur, low contrast, and color shift (see Figure 2.12). Most existing methods are optimized for a single type of degradation, reducing adaptability in multi-degraded scenarios.

Figure 2.12: *A graphical demonstration of multiple degraded images.*

- Most datasets and models focus on relatively clearer ocean environments, overlooking riverine turbidity (see Figure 2.13). Muddy water introduces extreme scattering, non-uniform blur, and severe visibility loss not captured in common benchmarks. Consequently, existing models generalize poorly to such conditions, failing to recover object boundaries and textures.

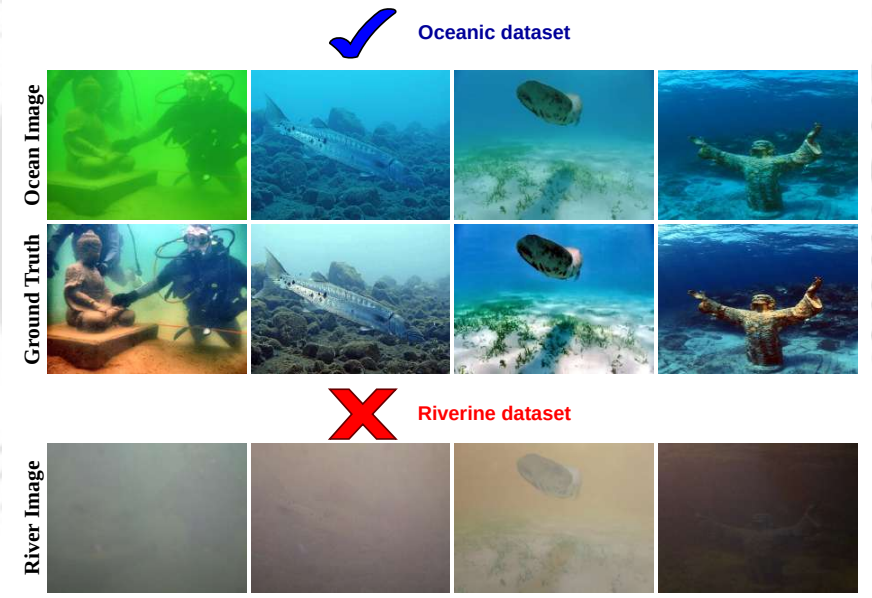


Figure 2.13: *A graphical demonstration of ocean and river images.*

2. With respect to underwater image super-resolution, the current literature has the following limitations:

   - Although current UISR approaches improve spatial resolution, they often struggle to accurately reconstruct high-frequency details, such as thin structures and small objects, under severe underwater degradation conditions. This often leads to over-smoothed textures and color distortion (refer to Figure 2.14), which can

obscure object boundaries and reduce the reliability of downstream tasks, such as detection and segmentation.

Figure 2.14: *A graphical demonstration of the limitations of existing UISR methods.*

- Existing UISR methods use complex deep learning models for high-quality reconstruction. However, these models typically require a significant amount of computing power. This makes them impractical for real-time processing on devices with limited resources, such as AUVs and ROVs.

2.11 Evaluation Metrics

Processing a degraded underwater image using enhancement or super-resolution models can significantly enhance its visual quality. The quality of these images can be assessed using two primary evaluation strategies: qualitative and quantitative. Qualitative evaluation relies on human visual perception and subjective judgment, without the use of explicit numerical reference parameters. In contrast, quantitative evaluation is based on mathematical analyses utilizing well-defined numerical measures. Quantitative quality metrics can also be categorized into two types: (1) full-reference metrics, which need a ground-truth image for comparison, and (2) reference-less metrics, which do not need ground-truth. In this work, several quantitative quality metrics are used to assess the performance of the proposed underwater image enhancement and super-resolution models, as discussed below:

2.11.1 Full-reference Metrics

Full-reference metrics assess the quality of an enhanced or super-resolved image by approximating it directly with a clean ground-truth image. These metrics need the actual reference

image to assess similarity.

- **Peak Signal-to-Noise Ratio (PSNR):** PSNR [113] is commonly used to measure image quality. It compares a ground-truth image (G) to an enhanced or super-resolved version (P), both of which are sized $a \times b$ pixels. PSNR is based on Mean Squared Error (MSE), which measures the difference between the pixels of the two images. A higher PSNR value indicates that an enhanced or super-resolved image appears more similar to the original, suggesting improved visual quality.

  Mathematically, PSNR is defined as:

$$\text{PSNR}(G, P) = 10 \log_{10} \left(\frac{peak^2}{\text{MSE}(G, P)} \right), \tag{2.13}$$

  where *peak* denotes the maximum possible pixel intensity value. For a x-bit image,

$$\text{peak} = 2^x - 1. \tag{2.14}$$

  The MSE between the original image G and the restored image P, both of size $a \times b$, is given by:

$$\text{MSE}(G, P) = \frac{1}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} (G_{ij} - P_{ij})^2. \tag{2.15}$$

- **SSIM:** SSIM [114] is a widely used full-reference image quality metric that evaluates the similarity between a ground-truth image G and its enhanced version P by considering luminance, contrast, and structural information. Unlike pixel-wise error metrics such as PSNR, SSIM is designed to better align with human visual perception by focusing on structural consistency between images.

  SSIM is computed by combining three comparison components: luminance comparison, contrast comparison, and structural comparison. The overall SSIM index is defined as:

$$\text{SSIM}(G, P) = l(G, P) \cdot c(G, P) \cdot s(G, P), \tag{2.16}$$

where

$$l(G,P) = \frac{2\mu_G\mu_P + C_1}{\mu_G^2 + \mu_P^2 + C_1}, \tag{2.17}$$

$$c(G,P) = \frac{2\sigma_G\sigma_P + C_2}{\sigma_G^2 + \sigma_P^2 + C_2}, \tag{2.18}$$

$$s(G,P) = \frac{\sigma_{GP} + C_3}{\sigma_G\sigma_P + C_3}. \tag{2.19}$$

Here, μ_G and μ_P denote the mean intensities of images G and P, respectively, σ_G and σ_P represent the corresponding standard deviations, and σ_{GP} denotes the covariance between the two images. The constants C_1, C_2, and C_3 are small positive values introduced to avoid numerical instability.

The SSIM value typically ranges between 0 and 1, where a higher value indicates greater structural similarity and better perceptual image quality.

- **Multi-Scale Structural Similarity Index Measure (MS-SSIM):** MS-SSIM [115] is an extension of SSIM that evaluates image similarity across multiple spatial scales. By progressively downsampling the input images, MS-SSIM captures both fine-grained details at higher resolutions and coarse structural information at lower resolutions. This multi-scale analysis provides a more robust assessment of perceptual image quality.

  At each scale j, luminance, contrast, and structural comparisons are computed. The overall MS-SSIM score is obtained by combining these comparisons across M scales:

$$\text{MS-SSIM}(G,P) = \left[l_M(G,P)\right]^{\alpha_M} \prod_{j=1}^{M} \left[c_j(G,P)\right]^{\beta_j} \left[s_j(G,P)\right]^{\gamma_j}, \tag{2.20}$$

  where l_M is the luminance comparison at the coarsest scale, and c_j and s_j denote the contrast and structural comparisons at scale j. The exponents α_j, β_j, and γ_j control the relative importance of each component.

A higher MS-SSIM value indicates greater perceptual similarity between the restored image and the reference image.

- **Learned Perceptual Image Patch Similarity (LPIPS):** LPIPS [116] is a perceptual image quality metric that evaluates the similarity between a reference image G and a restored image P using deep feature representations extracted from pretrained convolutional neural networks. In LPIPS, networks such as AlexNet, VGG, or SqueezeNet are employed as feature extractors, with AlexNet being one of the most commonly used backbones.

  Both images are passed through the pretrained network, and the perceptual distance is computed by comparing the corresponding deep feature maps. The LPIPS metric is defined as:

$$\text{LPIPS}(G, P) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{f}_l^G(h,w) - \hat{f}_l^P(h,w) \right) \right\|_2^2, \tag{2.21}$$

  where $\hat{f}_l^G$ and $\hat{f}_l^P$ denote the normalized feature maps extracted from layer l of AlexNet for images G and P, respectively. The term w_l represents learned channel-wise weights, while H_l and W_l denote the spatial dimensions of the feature maps.

  Lower LPIPS values indicate greater perceptual similarity between the restored image and the reference image.

2.11.2 Reference-less metrics

These objective metrics assess image quality without depending on corresponding clean reference images, making them appropriate for real-world scenarios where ground truth is unavailable.

- **Underwater Image Quality Measure (UIQM):** UIQM [117] is specifically designed to evaluate the visual quality of underwater images. Unlike full-reference metrics, UIQM does not require a ground-truth image and instead assesses image quality

based on perceptually relevant characteristics affected by underwater degradation.

UIQM is computed as a weighted combination of three components: colorfulness, sharpness, and contrast, and is defined as:

$$\text{UIQM} = c_1 \cdot \text{UICM} + c_2 \cdot \text{UISM} + c_3 \cdot \text{UIConM}, \tag{2.22}$$

where UICM measures colorfulness, UISM evaluates image sharpness, and UIConM quantifies contrast. The coefficients c_1, c_2, and c_3 are empirically determined weights.

A higher UIQM value indicates better perceptual quality of the underwater image.

- **Underwater Image Sharpness Measure (UISM):** UISM [117] evaluates the sharpness of an underwater image by measuring the strength of edge information. UISM is designed to capture the loss of fine details and edge blurring caused by underwater scattering and absorption.

  UISM is computed using edge detection operators, such as the Sobel filter, followed by statistical aggregation of edge intensities across the image. Stronger and well-defined edges result in higher UISM values, indicating improved sharpness and detail preservation.

  Higher UISM scores correspond to sharper images with better visibility of fine structures.

- **Blind/Reference-less Image Spatial Quality Evaluator (BRISQUE):** BRISQUE [118] evaluates perceptual image quality based on Natural Scene Statistics (NSS). BRISQUE operates directly in the spatial domain and measures deviations from statistical regularities observed in high-quality natural images.

  The metric extracts locally normalized luminance features and models their distribution using a trained regression model to predict image quality scores. Unlike underwater-specific metrics, BRISQUE is content-agnostic and can be applied to a wide range of image types.

Lower BRISQUE scores indicate better perceptual image quality.

2.12 Datasets

In this dissertation, various benchmarks have been utilized to perform comprehensive experiments against the SOTA works.

1. **Underwater image enhancment:**

   - **EUVP:** The EUVP [32] dataset contains 11,345 image pairs for underwater enhancement tasks. The testing set consists of 515 paired images. The size of each image is 256×256.
   - **UIEB:** The UIEB [30] dataset, comprising 890 image pairs, was utilized for the experiments. Following the method in [7], 800 images were randomly chosen for training, and 90 images were set aside for testing. All images were resized to 256×256 pixels. Additionally, 60 challenge images from the dataset were used to evaluate the model without ground truth references.
   - **SUIM-E:** The SUIM-E [75] dataset consists of 1,635 images, with 1,525 images used for training and the remaining 110 images reserved for testing. The image size is standardized to 256×256 pixels for both training and testing.
   - **LSUI:** The LSUI [74] dataset consists of 4279 image pairs, each with a size of 256×256, for the UIE task. Following the method of [75], 3879 images are randomly selected for training the model, while the remaining 400 images are dedicated for testing.

2. **Underwater image super-resolution:**

   - **UFO-120:** The UFO-120 dataset was used for the underwater image super-resolution task, which consists of 1,620 image pairs. Out of these, 1,500 images were utilized for training and 120 images for testing. The dataset includes degraded LR images with dimensions of 320×240 and their corresponding HR

images at 640 × 480, specifically for 2× SR. To achieve 3× and 4× SR, the degraded LR images were downsampled.

- **USR-248:** The USR-248 dataset is used for UISR task. It contains 1060 image pairs for training and 248 image pairs for testing. The degraded LR images of size 320 × 240, 160 × 120, and 80 × 60 are presented for 2×, 4×, and 8× super-resolution, respectively. The size of each HR image is 640 × 480.

2.13 Summary

This chapter begins with a background section that provides an overview of CNN, Transformers, and SSM. It then discusses key concepts, including the 2D DWT, the FFT, and the DoG. Following this, a comprehensive survey of existing approaches for UIE and UISR tasks is presented. The limitations observed in current research are also highlighted to identify areas that require further exploration. Additionally, evaluation metrics used to assess the methods, along with a detailed description of the datasets used for training and evaluating SOTA methods, are provided. With this background and related literature in mind, the next chapter will present the first contribution of this thesis, which focuses on color channel-aware feature learning for ocean-water image enhancement.

"Take up one idea. Make that one idea your life-think of it, dream of it, live on that idea."

~Swami Vivekananda

3

Color Channel-aware Feature Learning for Ocean-water Image Enhancement

Underwater image enhancement is essential to mitigate the environment-centric noise in images, such as haziness, color degradation, etc. With most existing works focused on processing an RGB image as a whole, the explicit context that can be mined from each color channel separately goes unaccounted for, ignoring the effects produced by the wavelength of light in underwater conditions. With this in mind, this chapter discusses the two works for the task of enhancing ocean-water images.

3.1 Lit-Net: Harnessing Multi-resolution and Multi-scale Attention for Underwater Image Enhancement

In recent research, deep learning-based methods have exhibited remarkable outcomes when applied to low level vision tasks. It has been noticed that a lot of State-of-the-art Underwater Image Restoration (UIR) works process the R-G-B channels with the same receptive field sizes. However, the same receptive fields for each R-G-B channel may not be suitable in most underwater situations. Sharma et al. [72] have shown that different sizes of receptive fields help in UIR. Still, it needs to capture the desired global effects that require considering the relationships between color components. Deep neural networks designed for UIR are broadly categorized into the single scale (high-resolution) [30,72,109] or an encoder-decoder feature processing [31,32,119]. High-resolution (single-scale) networks avoid down-sampling, resulting in images with more spatially precise information. However, these networks become less beneficial at encoding semantic information. For encoder-decoder based approaches, the encoder gradually maps the input to a low-resolution representation, and the decoder reverses the mapping to obtain the original resolution. These strategies grasp a broad context by reducing spatial resolution; however, finer spatial details are lost, making image reconstruction harder.

Underwater image enhancement is a technique that requires pixel-to-pixel correspondence between the input and output images and is sensitive to the position of pixels. Knowing that the underwater images are mainly bluish and greenish, removing only the undesired degraded content is crucial while preserving all the essential spatial information, including true edges and texture. For this, different sizes of receptive fields are used for different color channels (such as R (3×3), G (5×5), and B (7×7)) in the proposed architecture. This enables the model to perform multi-resolution analysis across different color channels, preserving spatial (high-resolution) information and capturing both local and global context. Additionally, the entire RGB image is processed with a 1×1 convolution to capture the

desired global effects, considering the relationships between color components. However, a 1×1 convolution outputs an image with the same width and height as the input. It acts as both a channel collapse and a color messenger to the preceding layers [120]. A new multi-scale strategy is incorporated to maintain semantic (encoder-decoder) information. Apart from that, the proposed work uses the 1x1 convolution layer in the encoder network to speed up the number of operations. Furthermore, the incorporation of attention-based skip connections leads to more discriminative representations.

Figure 3.1: *Visual comparisons of enhancement example on low contrast underwater image. The proposed Lit-Net model enhanced the contrast and removed the color deviation.*

Fig. 3.1 visually compares the proposed scheme and recent best-published works on the low-contrast real underwater image. It is observed that the proposed Lit-Net method produced visually more accepted results than recent best-published works. It is observed that most of the existing methods generate a greenish tone in their resulting image, while Lit-Net yields a visually improved color-corrected result. The major contributions of this work are outlined as follows:

- Propose a lightweight multi-stage network for UIE. The first stage is used for multi-resolution image analysis while retaining its original resolution. The second stage enhances the intermediate features while retaining the important information. The last stage focuses on the reconstruction of the enhanced and super-resolved image.

- Design a novel encoder block where each encoder layer has four parallel 1x1 convolution layers to pay attention to the dominant color and capture the local information precisely. It also speeds up the number of operations as it is less expensive than the larger-size of filters.

- Modified the l_1 loss function to a weighted color channel specific l_1 loss (cl_1) function. It helps to recover the color and detail information.

This work provides extensive experiments against the best-published SOTA in UIE to show the benefit. The performance of the proposed model has been demonstrated by comparing it with existing methods through various tasks, including underwater semantic segmentation and object detection.

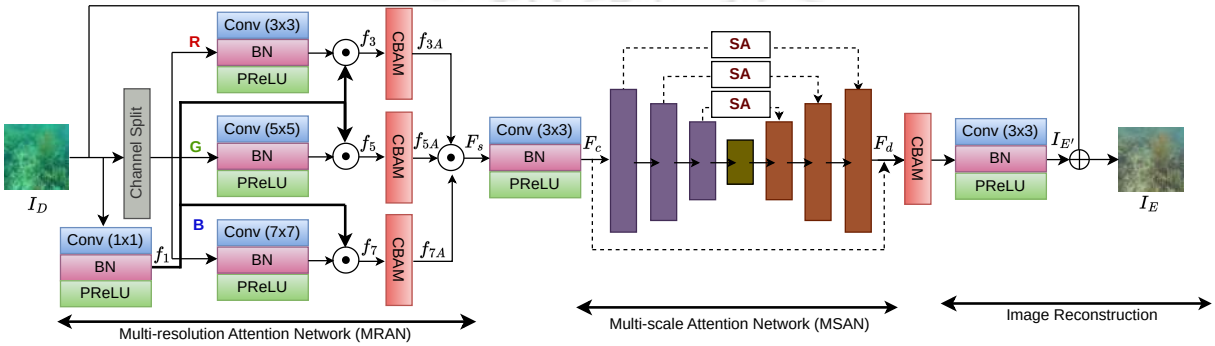


Figure 3.2: *Overview of the proposed model for UIE. The model accepts a degraded underwater image as input and produces an improved image that is visually and spatially enhanced.*

3.1.1 Proposed Method

3.1.1.1 Network Architecture

In this work, a deep learning-based architecture is proposed to enhance underwater images. The overall model is depicted in Fig. 3.2. The model is divided into three distinct modules as follows:

- Multi-resolution Attention Network (MRAN)
- Multi-scale Attention Network (MSAN)
- Image Reconstruction

The MRAN extracts shallow features that correspond to low-frequency details from the input degraded underwater image (I_D). The purpose of MRAN is to analyze the input image in a multi-resolution way while maintaining its original resolution. To obtain the most prominent

features, followed by the MRAN module, the proposed work designs a MSAN by extracting the rich features related to high-frequency details from the input image while retraining the important information through attention-based skipped connection. Finally, an image reconstruction module is devised to restore the enhanced and super-resolved image.

3.1.1.2 Multi-resolution Attention Network

In the proposed MRAN, full-resolution input images are utilized to develop multi-resolution feature representations. This is achieved by employing different kernel sizes, which provide varying receptive fields to analyze both global and local information. The MRAN consists of four convolutional layers, each using distinct kernel sizes of 1×1, 3×3, 5×5, and 7×7, along with batch normalization and PReLU activation. Channel-specific features for the R, G, and B channels are generated using 3×3, 5×5, and 7×7 kernels, while features from the entire RGB image are extracted with a 1×1 kernel. The RGB features are concatenated with the channel-specific features. These concatenated features pass through attention blocks, and the final output of the MRAN block is obtained by concatenating the results. Shallow contextual features (F_s) are derived from the input (I_D).

$$F_s = f_{3A} \odot f_{5A} \odot f_{7A}, \tag{3.1}$$

where, $\odot$ indicates the concatenation operation and the attentive contextual features (f_{kA}) with receptive field size k can be estimated as

$$\begin{aligned} f_{3A} &= A_n(f_3) \\ f_{5A} &= A_n(f_5) \\ f_{7A} &= A_n(f_7) \end{aligned} \tag{3.2}$$

The Convolutional Block Attention Module (CBAM) [1] module is used as an attention block (A_n) to refine the obtained features f_k. The architecture of CBAM is shown in Fig. 3.3. CBAM inputs an intermediate feature map to retrieve the channel and spatial attention

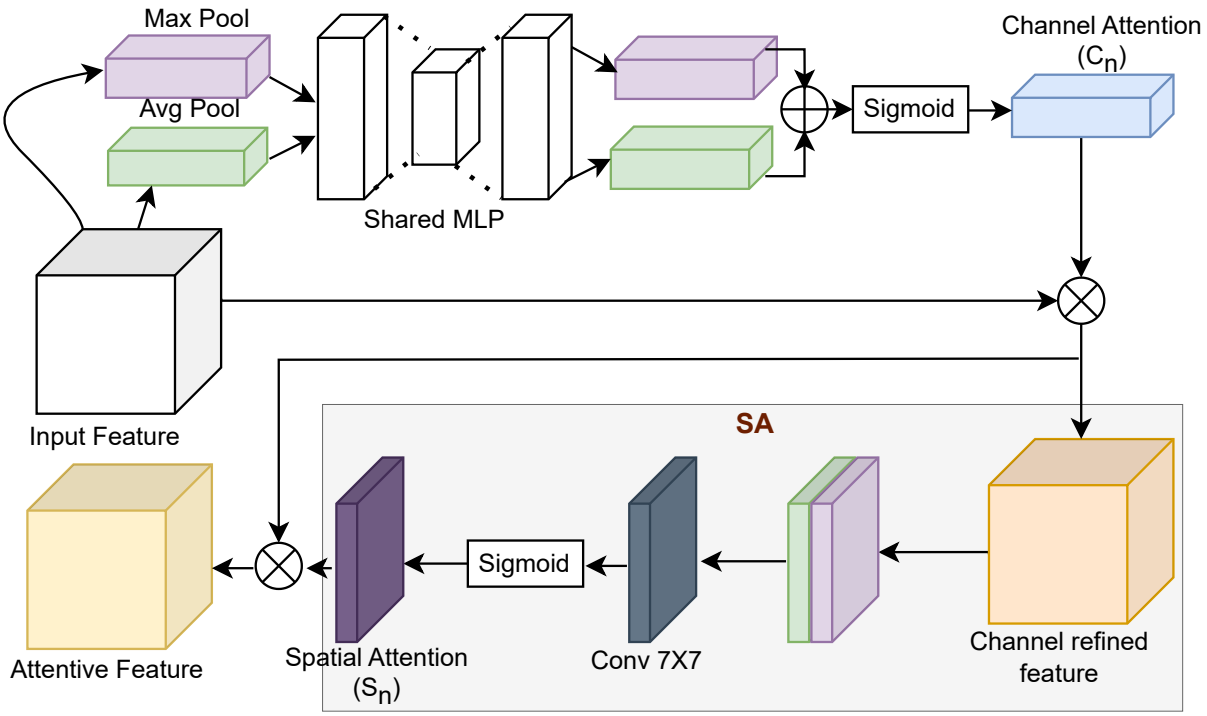


Figure 3.3: *Network architecture of CBAM block [1].*

features. The purpose of channel attention is to focus on identifying the important features present in an input image. It aims to determine 'what' aspects of the image are significant in terms of representing the desired information. Spatial attention concentrates on 'where' is an informative part. The refine features f_k can be computed as

$$\begin{aligned} f_1 &= p_r(b_n(c^{1\times1}(I_D))) \\ f_3 &= p_r(b_n(c^{3\times3}(I_D^R))) \odot f_1 \\ f_5 &= p_r(b_n(c^{5\times5}(I_D^G))) \odot f_1 \\ f_7 &= p_r(b_n(c^{7\times7}(I_D^B))) \odot f_1 \end{aligned} \tag{3.3}$$

where $c^{k\times k}$, b_n, and p_r represent a convolution operation, batch normalization [121] and parametric ReLU layers, respectively. $k \times k$ indicates the kernel size. I_D^C $(C \in R, G, B)$ indicates the individual channel of a degraded image.

The output generated by MRAN is first fed into a convolution layer, which is then used

as input for MSAN.

$$F_c = p_r(b_n(c^{3\times3}(F_s))) \tag{3.4}$$

3.1.1.3 Multi-scale Attention Network

The proposed MSAN module is a UNet-like architecture where each encoder and corresponding decoder layer is linked with an attention-based skipped connection. In this work, a novel extension (refer to Fig. 5.5) is used in each encoder layer to analyze the different input resolutions more accurately. To do this, this work maps the high-resolution feature (F_c) to a low-resolution representation and slowly retrieves the high-resolution representation from the lower level. Generally, the encoder-decoder process learns semantically-richer feature representation (f_{ds}) (multi-scale contextual information), but it is not spatially precise due to down-sampling. To get both semantically and spatially richer features (F_d), the features map f_{ds} is concatenated with the high-resolution features map (F_c). Also, to ensure the feature re-usability, a spatial attention-based skip connection is used, as in Fig. 3.2. The output of MSAN can be estimated as

$$F_d = F_c \odot f_{ds} \tag{3.5}$$

The proposed work modified each encoder layer (refer to Fig. 5.5). Generally, a normal encoder layer uses a 3×3 convolutional layer and a maxpool layer which is computationally costly because of 3×3 kernel. Here, each encoder layer contains four 1×1 convolution layers in parallel. Finally, the four convoluted features are concatenated and passed them through a pixel unshuffle layer. Instead of a maxpool layer, a pixel-unshuffle is used to downsample each feature to minimize the information loss. Three encoder layers are used in MSAN.

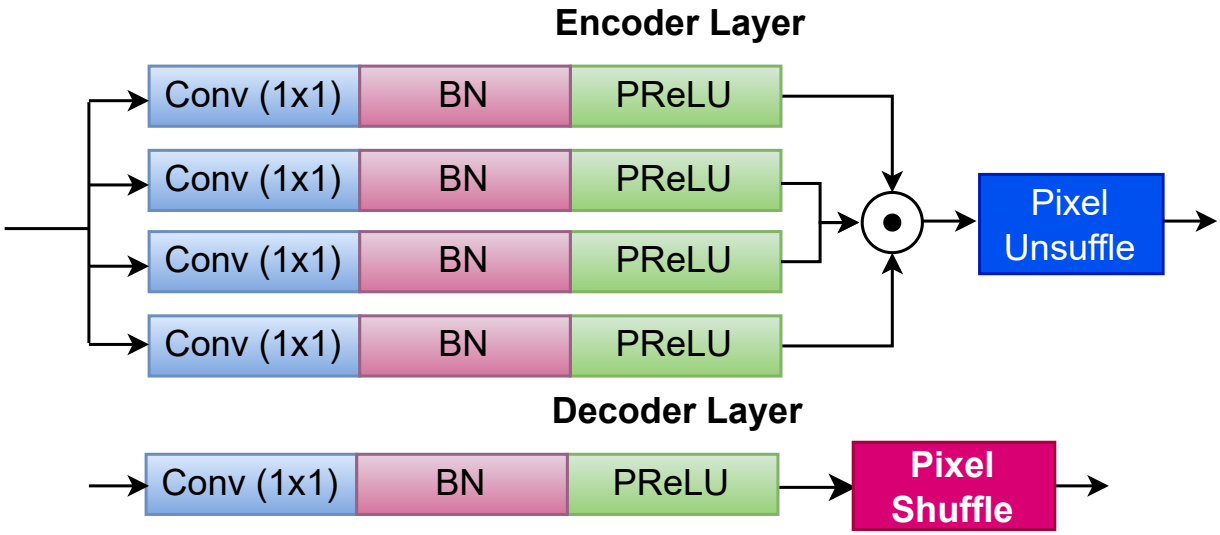


Figure 3.4: *Network architecture of proposed encoder and decoder layer.*

Mathematically, the encoder is formulated as

$$\begin{aligned} E_1^1 &= p_r(b_n(c^{1\times1}(F_c))) \\ E_1 &= P_{unshuffle}(E_1^1 \odot E_1^1 \odot E_1^1 \odot E_1^1) \\ E_2^1 &= p_r(b_n(c^{1\times1}(E_1))) \\ E_2 &= P_{unshuffle}(E_2^1 \odot E_2^1 \odot E_2^1 \odot E_2^1) \\ E_3^1 &= p_r(b_n(c^{1\times1}(E_2))) \\ E_3 &= P_{unshuffle}(E_3^1 \odot E_3^1 \odot E_3^1 \odot E_3^1) \end{aligned} \tag{3.6}$$

where E_i^1 represents the each branch of each encoder layer E_i $(i = 1, 2, 3.)$ and $P_{unshuffle}$ indicates the pixel-unshuffle operation. The output of the encoder can be computed as

$$B_{kn} = p_r(b_n(c^{1\times1}(E_3))) \tag{3.7}$$

Let S_n represent the spatial attention module and $P_{shuffle}$ represents the pixel shuffle operation as an up-sampling operation to increase the spatial resolution. The decoder can be

defined as

$$\begin{aligned} D_1 &= P_{shuffle}(B_{kn}) \odot S_n(E_3) \\ D_2 &= P_{shuffle}(D_1) \odot S_n(E_2) \\ f_{ds} = D_3 &= P_{shuffle}(D_2) \odot S_n(E_1) \end{aligned} \tag{3.8}$$

where, D_i $(i = 1, 2, 3.)$ indicates the each decoder layer.

3.1.1.4 Image Reconstruction

The final stage of the proposed model is used as the reconstruction module. In the case of underwater image enhancement, it takes the output features of MSAN (F_d) as input and feeds them into a convolutional layer to produce residual image $(I_{E'})$. Finally, the enhanced image is obtained by $I_E = (I_{E'}) + I_D$, where

$$I_{E'} = p_r(b_n(c^{3\times3}(F_d))) \tag{3.9}$$

3.1.1.5 Loss Function

Weighted Color channel specific L1 loss. It is observed in the literature that the l_2 loss function can lead to the appearance of blurry artifacts in the resulting enhanced images. Due to this, this work has chosen to employ l_1 loss rather than l_2 loss. To train Lit-Net, instead of using simple l_1 loss, a weighted color channel-wise l_1 loss is used for reconstructing high-quality underwater images. This is because underwater images can be affected differently by the water medium in each channel, and thus, the cl_1 loss enables the model to focus more on the channels that are dominated in underwater. The weighted channel-specific l_1 loss can be expressed as follows:

$$cl_1 = w_r l_{1_r} + w_g l_{1_g} + w_b l_{1_b} \tag{3.10}$$

where, l_{1_r}, l_{1_g}, and l_{1_b} indicates the red (r), green (g), and blue (b) channel specific l_1 loss, these can be formulated as

$$\begin{aligned} l_{1_r} &= \frac{1}{n}\sum_{i=1}^{n} ||ES(I_{D_i})_r - (I_{O_i})_r||_1 \\ l_{1_g} &= \frac{1}{n}\sum_{i=1}^{n} ||ES(I_{D_i})_g - (I_{O_i})_g||_1 \\ l_{1_b} &= \frac{1}{n}\sum_{i=1}^{n} ||ES(I_{D_i})_b - (I_{O_i})_b||_1 \end{aligned} \tag{3.11}$$

where ES(.) represents the process of Lit-Net, I_O is the original clean image. Empirically, the values of w_r, w_g, and w_b are set to 1, 1.5, and 2, respectively.

Perceptual loss. Although training with pixel loss enhances PSNR, it does not improve perceptual quality. To address this flaw, this work integrated the Perceptual loss [122] to make it more perceptually improved. This loss function helps to ensure that the enhanced image maintains its fine details and textures. To accomplish this goal, a VGG16 ($V_f(\phi)$) [123] model is pre-trained on a large-scale image database (ImageNet). The cost function is defined as the Euclidean distance between the relu2_2 features obtained from the predicted image and the actual clean image, and it can be expressed as follows:

$$l_p = \frac{1}{n}\sum_{i=1}^{n} ||V_f(ES(I_{D_i});\phi) - V_f(I_{O_i};\phi)||_2^2 \tag{3.12}$$

SSIM loss. Besides using l_1 and perceptual losses, to minimize the structural dissimilarities between I_E and I_O, the SSIM loss [114] is included. The SSIM is a way to quantify how similar two images are in terms of their structure, luminance, and contrast and can be expressed as:

$$SSIM(c) = \frac{2.\mu_a.\mu_b + K_1}{\mu_a^2 + \mu_b^2 + K_1}.\frac{2.\sigma_{ab} + K_2}{\sigma_a^2 + \sigma_b^2 + K_2} \tag{3.13}$$

where a,b are patches from I_E or I_{SR}, and I_O, respectively, μ, σ, and σ_{ab} denote mean, standard deviation and covariance, given c the center of patches. K_1 and K_2 are fixed

parameters. The SSIM loss can be expressed as follows:

$$l_s = \frac{1}{2n}\sum_{i=1}^{n} 1 - SSIM(ES(I_{D_i}), I_{O_i}) \tag{3.14}$$

Finally, the proposed method is trained by minimizing the following losses:

$$\mathcal{L}_T = \lambda_1 cl_1 + \lambda_p l_p + \lambda_s l_s \tag{3.15}$$

where λ_1, λ_p, and λ_s have been experimentally set to 1, 0.02, and 0.5, respectively.

3.1.2 Experiments

In this section, the proposed work performs experiments on Four popular standard datasets (EUVP, UIEB, SUIM-E, and LSUI) to show the efficacy of Lit-Net model. Apart from the datasets mentioned above, this work also used the underwater color cast set (**UCCS**) [124], **U45** [125], and UIDEF [126] dataset for testing. The performance of Lit-Net model is evaluated against state-of-the-art approaches and conduct an ablation study to assess the impact of each component of Lit-Net.

Table 3.1: *MSE, PSNR, SSIM, MS-SSIM, LPIPS, UIQM, UISM, and BRISQUE comparison on EUVP dataset with the best-published works for UIE. First, second, and third best performance are represented in red, blue, and green colors, respectively. ↓ indicates lower is better.*

Method	MSE↓	PSNR	SSIM	UIQM	LPIPS↓	UISM	MS-SSIM	BRISQUE↓
UGAN [31]	0.355	26.551	0.807	2.896	0.220	6.833	0.936	35.859
UGAN-P [31]	0.347	26.549	0.805	2.931	0.223	6.816	0.935	35.099
FUnIE-GAN [32]	0.386	26.220	0.792	2.971	0.212	6.892	0.924	30.912
FUnIE-GAN-UP	0.600	25.224	0.788	2.935	0.246	6.853	0.925	34.070
Deep SESR [109]	0.325	27.081	0.803	3.099	0.206	7.051	0.940	35.179
DWNet [72]	0.276	28.654	0.835	3.042	0.173	7.051	0.950	30.856
Ushape [74]	0.370	26.822	0.811	3.052	0.187	6.843	0.949	35.648
Proposed	0.225	29.477	0.851	3.027	0.169	7.011	0.954	32.109

3.1.2.1 Experimental Setup

Training Process. The proposed Lit-Net model is implemented based on an open-source deep learning framework called Pytorch. In the training phase, the Adam optimizer is utilized with an initial learning rate of $2e-4$. All the experiments have been conducted on a Ubuntu 20.04 LTS operating system and one Nvidia A100 GPU. During the training process, a batch size of 5 was employed for the model.

Evaluation Metric. To ensure a fair comparison between the proposed model and existing state-of-the-art approaches, this work assessed the performance of the model using widely used reference and reference-less image quality metrics. These include MSE, PSNR, SSIM, MS-SSIM, LPIPS, UIQM [117], UISM, and BRISQUE.

State-of-the-Art Methods. In the objective of UIE, the proposed work is compared with several existing SOTA studies, which include: UDCP [127] ($ICCVW-13$), GBdehaze [49] ($ICASSP-16$), IBLA [50] ($TIP-17$), ULAP [56] ($PCM-18$), CBF [57] ($TIP-18$), GDCP [128] ($TIP-18$) UGAN [31] ($ICRA-18$), WaterNet [30] ($TIP-20$), UWCNN [18] ($PR-20$), FUnIE GAN [32] ($RAL-20$), DeepSESR [109] ($RSS-20$), UIEC^ 2 [129] ($SPIC-21$), Ucolor [33] $TIP-21$, UWCNN-SD [21] ($IJOE-21$) MILE [37] ($TIP-22$), PUIE [130] ($ECCV-22$), SGUIE-Net [75] ($TIP-22$), UGIF-Net [131] ($TGRS-23$), DWNet [72] ($TOMM-23$), Ushape [74] ($TIP-23$), DM-Underwater [132] ($MM-23$), and Wavelet [79] ($CVPR-24$).

The proposed method employs identical training and test sets as other deep learning-based approaches. For a fair comparison, the training and testing image resolution was the same for all the methods as ours. Apart from that, the same environment (Python/Matlab) is used for calculating the measurement metrics.

3.1.2.2 Comparisons with State-of-the-Arts

Both quantitative and qualitative comparison results of the proposed model with the recent best-published methods are presented in the following subsections:

Quantitative Evaluation. This sub-section has focused on presenting a quantitative

Table 3.2: *PSNR, SSIM, MS-SSIM, LPIPS, UIQM, UISM, and BRISQUE comparison on UIEB test set with the best-published works for UIE. First, second, and third best performances are represented in red, blue, and green colors, respectively. ↓ indicates lower is better.*

Method	PSNR	SSIM	MS-SSIM	LPIPS↓	UIQM	UISM	BRISQUE↓
UDCP [127]	13.026	0.545	0.769	0.283	1.922	7.424	24.133
GBdehaze [49]	15.378	0.671	0.777	0.309	2.520	7.444	23.929
IBLA [50]	19.316	0.690	0.855	0.233	2.108	7.427	23.710
ULAP [56]	19.863	0.724	0.865	0.256	2.328	7.362	25.113
CBF [57]	20.771	0.836	0.890	0.189	3.318	7.380	20.534
UGAN [31]	23.322	0.815	0.932	0.199	3.432	7.241	27.011
UGAN-P [31]	23.550	0.814	0.933	0.192	3.396	7.262	25.382
FUnIE-GAN [32]	21.043	0.785	0.890	0.173	3.250	7.202	24.522
SGUIE-Net [75]	23.496	0.853	0.926	0.136	3.004	7.362	24.607
DWNet [72]	23.165	0.843	0.929	0.162	2.897	7.089	24.863
Ushape [74]	21.084	0.744	0.895	0.220	3.161	7.183	24.128
Proposed	23.603	0.863	0.935	0.130	3.145	7.396	23.038

comparison between the proposed method and other existing SOTA approaches with respect to reference and no-reference quality evaluations. The results of these evaluations have been provided in Tables 4.1, 5.1, 4.4, 4.3, 3.5, 3.6, and 3.7. Table 4.1, 5.1, 4.4, 4.3, and 3.5 provide a quantitative comparison of UIE methods. Table 4.1 illustrates that the proposed approach has achieved better results than previous methods for full reference-based quality metrics on the EUVP test set. Besides, Lit-Net also achieves the second and third-best results in terms of non-reference measurement. It is observed that the proposed model has gained improvement on both SSIM and PSNR, 1.92% and 2.87%, respectively. Lit-Net has a low LPIPS score, indicating that the image patches are perceptually more similar to the ground truth. The Lit-Net has surpassed the performance of previous state-of-the-art methods in terms of PSNR, SSIM, and MS-SSIM on the UIEB test dataset, as demonstrated in Table 5.1. Additionally, the results also indicate that the proposed model is competitive with the recent schemes in terms of LPIPS and BRISQUE. Table 4.4 shows that Lit-Net achieves 2.67% SSIM gains over the previous best methods on the SUIM-E dataset. Proposed method provides a substantial gain of 0.53% MS-SSIM compared to SGUIE-Net. The corresponding LPIPS score is less, which indicates better perceptual quality. Table 4.3 demonstrates the UISM, UIQM, and BRISQUE scores of different methods on the UIEB challenge set. The table indicates that the proposed Lit-Net achieves the highest UISM value, second highest

Table 3.3: *PSNR, SSIM, MS-SSIM, LPIPS, UIQM, UISM, and BRISQUE comparison on SUIM-E test set with the best-published works for UIE. First, second, and third best performances are represented in red, blue, and green colors, respectively. ↓ indicates lower is better*

Method	PSNR	SSIM	MS-SSIM	LPIPS↓	UIQM	UISM	BRISQUE↓
UDCP [127]	12.074	0.513	0.742	0.270	1.648	7.537	22.788
GBdehaze [49]	14.339	0.599	0.743	0.355	2.255	7.400	20.175
IBLA [50]	18.024	0.685	0.849	0.209	1.826	7.341	20.957
ULAP [56]	19.148	0.744	0.871	0.231	2.115	7.475	21.250
CBF [57]	20.395	0.834	0.884	0.194	3.003	7.360	21.115
UGAN [31]	24.704	0.826	0.941	0.190	2.894	7.175	20.288
UGAN-P [31]	25.050	0.827	0.943	0.188	2.901	7.184	18.768
FUnIE-GAN [32]	23.590	0.825	0.913	0.189	2.918	7.121	22.560
SGUIE-Net [75]	25.987	0.857	0.945	0.153	2.637	7.090	25.927
DWNet [72]	24.850	0.861	0.941	0.133	2.707	7.381	20.757
Ushape [74]	22.647	0.783	0.917	0.213	2.873	7.061	22.876
Proposed	25.117	0.884	0.950	0.118	2.918	7.368	19.602

Table 3.4: *UISM, UIQM, and BRISQUE comparison on UIEB challenge set with the best-published works. First, second, and third best performances are represented in red, blue, and green colors, respectively.*

	UDCP [127]	GBdehaze [49]	IBLA [50]	ULAP [56]	CBF [57]	UGAN [31]	FUnIE-GAN [32]	Ucolor [33]	SGUIE-Net [75]	DWNet [72]	Ushape [74]	Proposed
UISM	7.358	7.399	7.310	7.239	7.396	7.158	7.140	7.244	7.269	6.677	7.292	7.458
UIQM	1.566	2.280	2.142	1.815	2.810	2.662	2.768	2.483	2.527	2.269	2.783	2.795
BRISQUE↓	29.658	26.264	24.972	30.472	29.213	25.118	24.773	25.093	27.320	31.160	23.616	24.755

BRISQUE value and UIQM value over the SOTA methods. Table 3.5 demonstrates the quantitative results of U45 and UCCS datasets. The proposed approach attains the top UIQM scores for the UCCS and U45 datasets, respectively. Table 3.6 shows the comparison with multi-scale method [126] on the UIDEF dataset, and LitNet is getting comparable results. In the LSUI dataset, LitNet achieves the best and second best performance, as is evident from Table 3.7. Lit-Net exhibits superior performance on various datasets, implying the model's robustness. Table 3.8 shows that the proposed model takes second less number of parameters and GFLOPs. However, this marginal increment also provides a relative performance boost of up to 2.87% and 1.92% regarding PSNR and SSIM over the SOTA on EUVP dataset, which justifies the trade-off between performance and parameters.

Qualitative Evaluation. In this subsection, this work presented the qualitative performance of the Lit-Net in the UIE task. Also, It has carried out the qualitative evaluation with the other existing works. The visual comparison of UIE is shown in Fig. 3.5, 3.6,

Table 3.5: *UIQM comparison on UCCS and U45 datasets with the best-published works. First, second, and third best performances are represented in red, blue, and green colors, respectively.*

UCCS	GDCP [128]	WaterNet [30]	FUnIE-GAN [32]	UWCNN [18]	UIEC^2 [129]	Ucolor [33]	UWCNN-SD [21]	MLIE [37]	PUIE [130]	UGIF-Net [131]	DWNet [72]
	2.716	3.073	3.074	2.981	2.900	2.983	3.130	2.863	2.943	3.112	3.089
	INSPIRATION [133]	DM-Underwater [132]	Wavelet [79]	Ours							
	2.983	3.113	-	3.167							
U45	GDCP [128]	WaterNet [30]	FUnIE-GAN [32]	UWCNN [18]	UIEC^2 [129]	Ucolor [33]	UWCNN-SD [21]	MLIE [37]	PUIE [130]	UGIF-Net [131]	DWNet [72]
	2.370	3.006	2.563	3.110	3.152	3.185	3.134	2.550	2.625	3.264	3.128
	INSPIRATION [133]	DM-Underwater [132]	Wavelet [79]	Proposed							
	3.167	3.252	3.181	3.279							

Table 3.6: *Comparison on UIDEF dataset*

Methods	NIQE ↓	UIQM	UCIQE	CCF
Multi-scale [126]	5.266 ± 1.383	3.645 ± 0.647	0.656 ± 0.028	28.137 ± 9.216
Proposed	5.242 ± 1.122	3.615 ± 0.190	0.638 ± 0.016	29.462 ± 5.056

3.7, and 3.8. The results displayed in these figures indicate that Lit-Net generates the most visually enhanced images with respect to ground truth. From Fig. 3.5, one can easily notice that the existing methods IBLA [50] and ULAP [56] fail to recover degraded images. The methods CBF [57] exhibit issues of color deviation in their enhanced images. UGAN [31] and FUnIE-GAN [32] suffer from low contrast. SGUIE-Net [75] provides a better result, but it has a bluishness in the produced image. It can be observed from the third row of Fig. 3.6 that images produced by the proposed model look more perceptually improved and close to the ground truth. This may be due to the fact that the proposed approach better balances visual quality and color correction. It is also observed from Fig. 3.7 that the proposed Lit-Net model is capable of handling enhancement effectively, unlike other methods that often result in under and over-saturation, artifacts or fail to correct color deviations. Similar results on the test set of the UIEB challenge are presented in Fig. 3.8 to show the efficacy of proposed approach. Lastly, it has tested the proposed model with the UIDEF dataset and Figure 3.9 shows the visual result. The proposed method produces more consistent enhancement over the SOTA schemes in terms of color correction, contrast enhancement, and detail improvement.

Table 3.7: *Comparison on LSUI dataset*

Methods	UIEWD [134]	UWCNN [135]	UIEC^2 [129]	WaterNet [30]	Ucolor [33]	Ushape [74]	DM-Underwater [132]	Wavelet [79]	Proposed
PSNR	15.43	18.24	20.86	19.73	22.91	24.16	27.65	27.26	29.69
SSIM	0.7802	0.8465	0.8867	0.8226	0.8902	0.9322	0.8867	0.9437	0.9364
LPIPS↓	0.3962	0.3450	0.1432	0.1678	0.123	0.1028	0.1138	0.1096	0.1087

Table 3.8: *Comparison with the model parameters and GFLOPs of the SOTA model with the image size* 256×256. *Lower is better.*

	WaterNet	UGAN	FUnIE-GAN	Ucolor	SGUIE-Net	DWNet	Ushape	DM-Underwater	Wavelet	Proposed
Parameters (M)	24.8	57.17	7.71	157.4	18.55	0.48	65.6	10	-	0.54
FLOPs (G)	193.7	18.3	10.7	443.9	123.5	18.2	66.2	-	94.1	17.8

3.1.3 Ablation Study

In this section, the contributions of various parameters are analyzed in the proposed method.

Table 3.9: *Effectiveness of various cost functions on three datasets: EUVP, UIEB, and SUIM-E, for the task of UIE*

	EUVP			UIEB			SUIM-E		
	cl_1	$cl_1 + \lambda_p l_p$	$cl_1 + \lambda_p l_p + \lambda_s l_s$	cl_1	$cl_1 + \lambda_p l_p$	$cl_1 + \lambda_p l_p + \lambda_s l_s$	cl_1	$cl_1 + \lambda_p l_p$	$cl_1 + \lambda_p l_p + \lambda_s l_s$
PSNR	29.340	29.095	29.477	22.97	23.593	23.603	24.721	24.899	25.326
SSIM	0.830	0.849	0.851	0.849	0.859	0.863	0.874	0.881	0.884
UIQM	2.980	3.008	3.027	3.118	3.010	3.145	2.819	2.885	2.938

Impact of loss functions. This work has performed a series of experiments to demonstrate the influence of various loss functions incorporated in proposed model. From Table 4.7, it is evident that l_p loss, when added with cl_1 loss, has improved the UIE performance. The intuitive reason behind such improvement may be the low-level features in the improved images may be benefited from the perceptual loss. Moreover, it is observed from Table that incorporating the SSIM-based loss function (l_s) has improved the corresponding PSNR and SSIM values noticeably.

Effect of Attention Module. In Table 3.10, the impact of attention modules is shown. This work carried out the experiments on all the datasets. Here, the results of the UIEB and SUIM-E are presented. From the first, second, and third rows of Table 3.10, it is observed that the performance of the proposed scheme is relatively low when the attention module is not included in the MRAN block, MSAN block, and the whole Lit-Net. Intuitively, the addition of the CBAM module helps the proposed Lit-Net model to effectively preserve and recover the texture details of the degraded images. As shown in Fig. 3.10(b), (c), and (d), the enhanced image without the attention module fails to correct the color, local and global

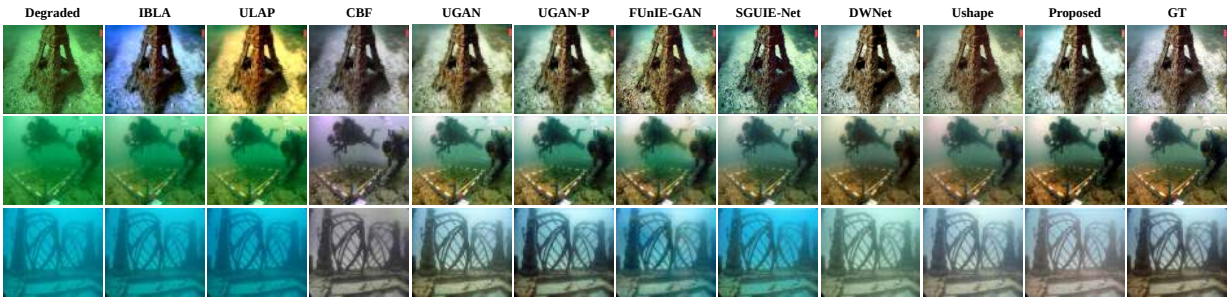


Figure 3.5: *Visual demonstration against the state-of-the-art UIE models on UIEB dataset.*

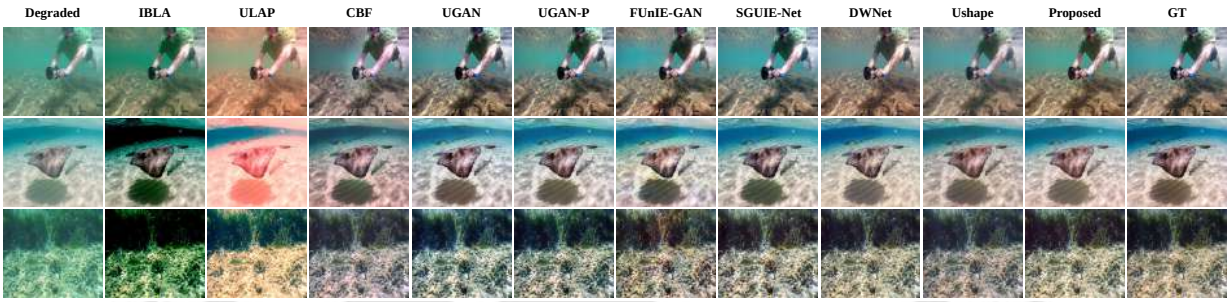


Figure 3.6: *Visual demonstration against the state-of-the-art UIE models on SUIM-E dataset.*

contrast.

Effect of different receptive fields and encoder block. To show the effect of receptive fields and encoder block, the experiments have been carried out with the following setup:

1. set 3×3 kernel size in the MRAN instead of three different kernels.

2. set one 1×1 kernel in the encoder of MSAN instead of four parallel 1×1.

The fourth row of Table 3.10 shows a substantial drop in SSIM, PSNR, and UIQM when this work performed the task mentioned above. Notably, Lit-Net with this setting has decreased the 0.15% and 0.98% PSNR value and 3.28% and 4.18% UIQM value on UIEB and SUIM-E datasets, respectively. Fig. 3.10(e) shows the visual result of the Lit-Net with these settings. It can be seen that the produced result fails to recover the color.

Effect of weighted color specific L1 loss and color channel split module. The weighted color specific l_1 loss is used to correct the color, which is affected by the wavelength of the underwater medium. The sixth row of Table 3.10 shows the result of proposed method

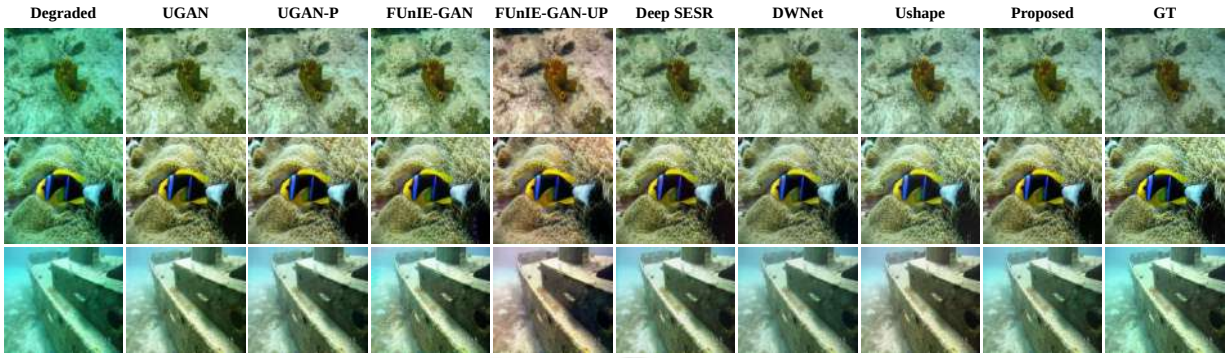


Figure 3.7: *Visual demonstration against the state-of-the-art UIE models on EUVP dataset.*

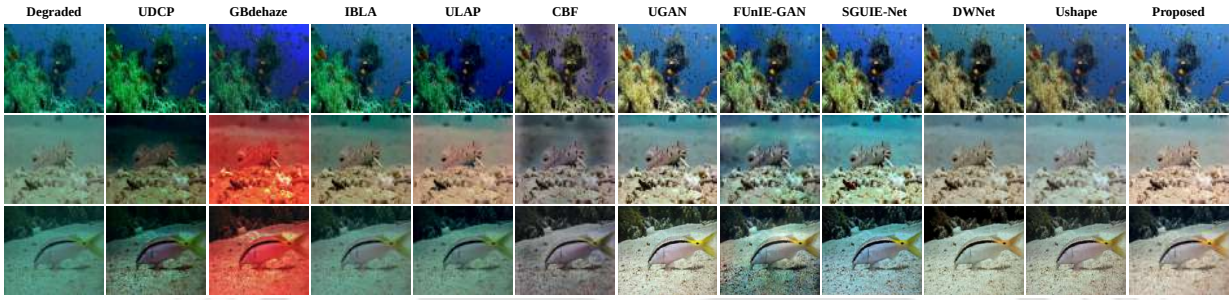


Figure 3.8: *Visual demonstration against the state-of-the-art UIE models on UIEB challenge test set.*

with l_1 loss along with other losses. It is observed that the proposed method with l_1 loss has decreased the PSNR, SSIM, and UIQM values as compared to cl_1 loss.

Similar to color-specific loss, Lit-Net processes the R, G, and B channels with different kernel sizes to capture the global context from the blue channel and the local context from the red channel. When the proposed method processes the entire RGB image instead of employing R, G, and B channel-wise processing, it can be observed from the $5th$ row of Table 3.10, Lit-Net provides a decrease of 1.086 dB and 0.495 dB PSNR. Compared with Lit-Net w/o-channel split, the full model improves the UIQM score with 2.29% and 2.91% on UIEB and SUIM-E test set, respectively. Fig. 3.10(f) shows the visual quality of Lit-Net w/o-channel split, it is less sharp and fails to recover the details. Fig. 3.10(g) shows that the full Lit-Net provides better color correction, detail recovery, and perceptual enhancement.

Effect of different kernels on different color channels. In the underwater scenario, the blue (B) color traverses the longest compared to the green (G) and red (R) color because of its shortest wavelength. This phenomenon causes most underwater images to appear blue.

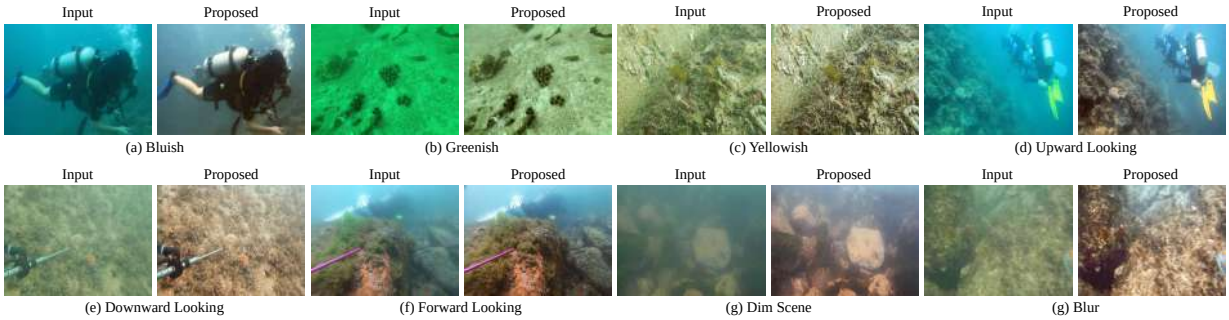


Figure 3.9: *Visual demonstration on UIDEF dataset.*

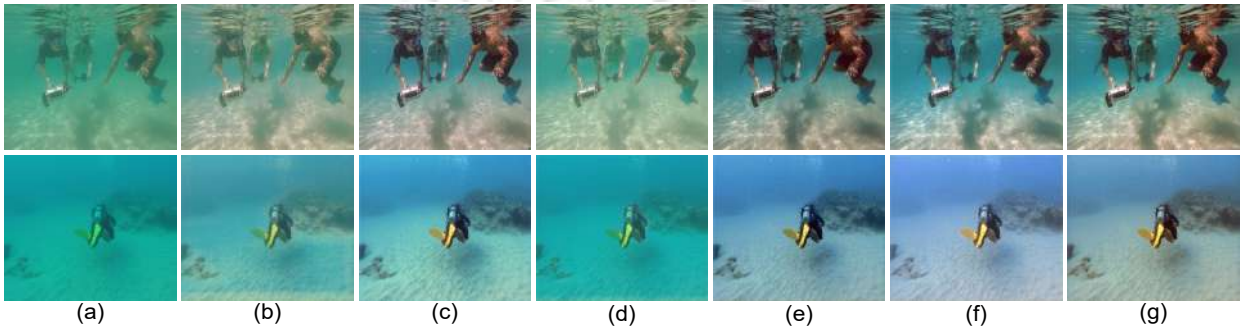


Figure 3.10: *Visual demonstration of different modules of Lit-Net. (a) Degraded image, (b) MRAN without attention module, (c) MSAN without spatial attention module, (d) Lit-Net without attention module, (e) Lit-Net with fixed size kernel, (f) Lit-Net without channel-wise split, and (g) Full Lit-Net*

The global coherence in a generic underwater image mostly aligns with the blue color. A larger receptive field effectively identifies global features, while a smaller one captures local details. Based on the traversing range, this work choose a 7×7 kernel for B, 5×5 for G, and 3×3 for R channel.

To verify the above observation proposed work tested the results using kernels of different sizes for different color channels. The result is shown in Table 3.11. From the table, it is evident that choosing a large kernel for blue and a small kernel for red improves the result.

3.1.3.1 Impact on High level Vision Applications

This work demonstrated the resilience of the improved images generated by the Lit-Net against state-of-the-art UIE strategies on several underwater high-level vision applications. Tasks like underwater object detection and single-image semantic segmentation have been taken into account for this. Also, the proposed work has extended the evaluation to in-

Table 3.10: *Effectiveness of various modules on two datasets: UIEB and SUIM-E for the task of UIE.*

Ablations			UIEB			SUIM-E		
	GFLOPS	Parameter(M)	PSNR	SSIM	UIQM	PSNR	SSIM	UIQM
MRAN w/o-Attention	17.852	0.5433	22.958	0.859	3.110	22.230	0.862	2.798
MSAN w/o-Attention	17.863	0.5443	23.255	0.860	3.113	24.499	0.884	2.856
Lit-Net w/o-Attention	17.837	0.5373	20.576	0.832	3.124	21.821	0.854	2.738
Lit-Net w fixed-size kernel	15.912	0.3199	23.568	0.855	3.042	24.871	0.882	2.796
Lit-Net w/o-channel split	14.587	0.4917	22.517	0.854	3.073	24.622	0.872	2.833
Lit-Net w L1 loss	17.871	0.5446	23.530	0.857	3.069	25.003	0.877	2.791
Proposed	17.871	0.5446	23.603	0.863	3.145	25.117	0.884	2.918

Table 3.11: *Ablation on UIEB dataset by interchanging the kernel sizes.*

Methods	PSNR	SSIM	UIQM
G = 3×3, R = 5×5, B = 7×7	22.931	0.854	3.088
B = 3×3, G = 5×5, R = 7×7	22.052	0.831	3.022
G = 7×7, B = 5×5, R = 3×3	23.127	0.858	3.113
R = 3×3, G = 5×5, B = 7×7	23.603	0.863	3.145

clude tests on foggy and low-contrast images following [136]. The results are presented in Figure 3.11. Proposed method shows strong performance on difficult low-contrast and foggy images. These images frequently exhibit poor visibility and a lack of detail, but the proposed approach effectively improves visual quality by enhancing contrast and clarity.

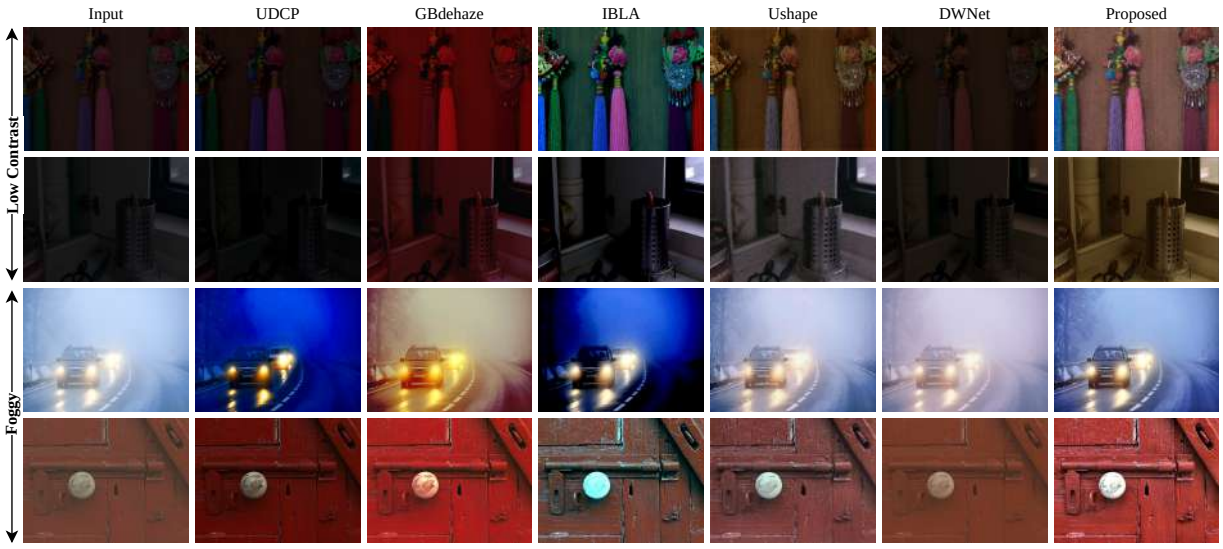


Figure 3.11: *Examples on foggy and low-contrast images.*

Underwater Semantic Segmentation. Semantic segmentation is a process of labelling or assigning a category to each pixel in an image. This technique is utilized to identify a group of pixels that collectively belong to separate categories for detailed scene understanding. The SUIM dataset [137] is used for underwater semantic segmentation.

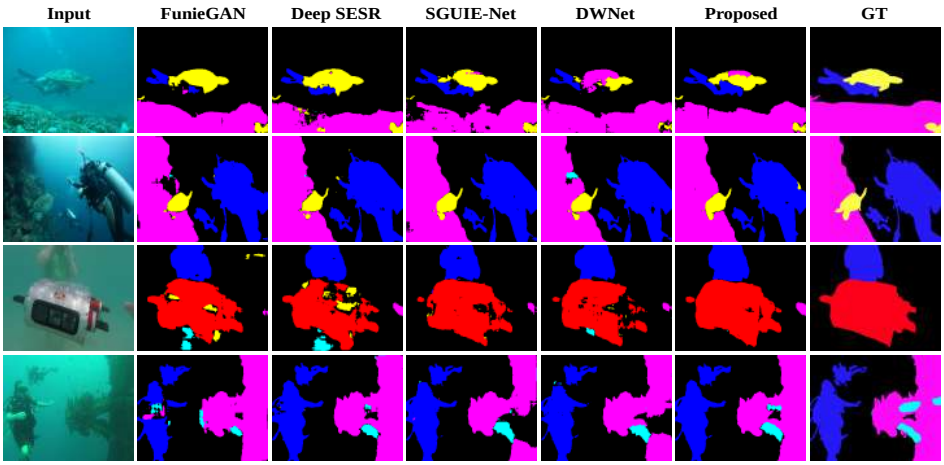


Figure 3.12: *Visual comparison of the semantic segmentation maps derived from the improved underwater images from four existing UIE works.*

First, this work enhances the images in the SUIM dataset using proposed Lit-Net. Second, it utilized the SUIM-Net [137] method for the semantic segmentation task to show the performance gain by using the enhanced images. Fig. 3.12 shows that generated segmentation maps are more refined than existing ones.

Table 3.12: *Mean average precision (mAP) comparison on detection task*

	Fish	Jellyfish	Penguin	Puffin	Shark	Starfish	Stingray	All
Degraded	0.815	0.929	0.749	0.602	0.749	0.838	0.781	0.780
FUNIE-GAN	0.744	0.923	0.654	0.445	0.664	0.716	0.755	0.700
Deep SESR	0.706	0.943	0.643	0.41	0.638	0.703	0.757	0.686
DWNet	0.839	0.948	0.752	0.551	0.757	0.819	0.806	0.782
Proposed	0.821	0.940	0.753	0.615	0.785	0.841	0.814	0.796

Underwater Object Detection. To validate the impact of UIE on object detection tasks, it utilizes the improved outcomes in a detection algorithm. For this task, it has used the YOLOv5 [138] and Aquarium dataset [139]. This work reported the quantitative results in Table 3.12. The results indicate that the proposed method exhibits better detection accuracy than the other competitive methods. Fig. 3.13 displays the visualization of detection results obtained on the Aquarium dataset. It is shown in Fig. 3.13 that in spite of visual enhancement, FUNIEGAN [32], Deep SESR [109], and DWNet [72] methods showed inferior results for the detection performance in comparison with proposed Lit-Net model.

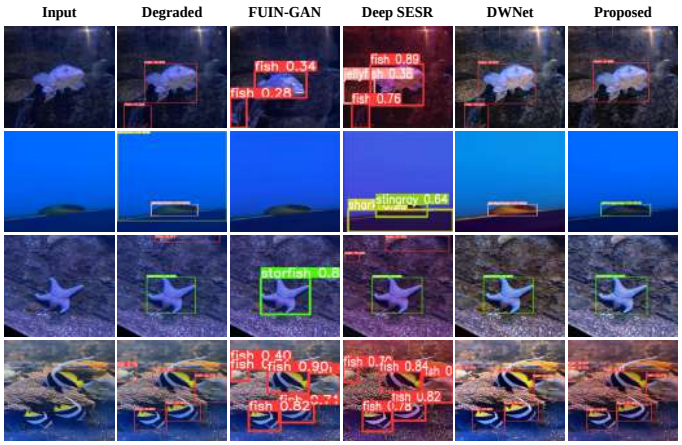


Figure 3.13: *Visual demonstration of object detection by utilizing the improved images from three existing UIE works.*

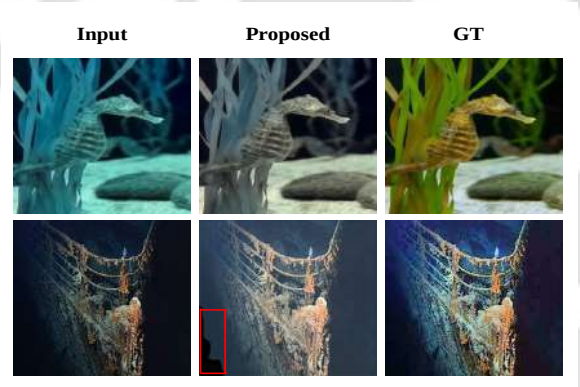


Figure 3.14: *Visual demonstration of object detection by utilizing the improved images from three existing UIE works.*

3.1.3.2 Failure Case

This work examines a few failure cases of Lit-Net, as depicted in Figure 3.14. Proposed approach needs to perform well in a few cases where the rate degradation is highly intense, such as extremely low contrast, irregular coloring, etc. For example, in the first image, the GT image contains color for the object and sea grass, and it is almost difficult to observe its color in the predicted image because it is tough to identify the colors from the original degraded image. As illustrated in the second image, the contrast of the lower left region is extremely diverse from the contrast of the lower region. Since proposed model lacks explicit provisions for extremely low-light image enhancement, a few resultant images may exhibit a tendency towards suboptimal enhancement. In the future, it can be introduced extremely

low-light correction modules to precisely enhance images with uneven illumination.

3.2 X-CAUNET: Cross-color Channel Attention with Underwater Image-enhancing Transformer

This work proposes a novel deep learning-based framework called **X-CAUNET (Cross-color channel attention with underwater image-enhancing transformer)** that looks at an underwater image from a finer perspective - mining information from individual color channels and expressing their inter-associativity towards learning enhancement elements for the entire image. Specifically, X-CAUNET splits an input image into red, green, and blue channels and extracts their local contextual visual features using convolutional receptive fields of multiple sizes. Since the blue color attenuates the slowest within an underwater environment due to shorter wavelength as compared to red and green, this work performs cross-color channel attention between the visual features of red-blue and green-blue channels using novel transformers. The resultant features are further passed through a Selective Kernel Feature Fusion (SKFF) block that uses fewer parameters than simple concatenation while generating better results. Additionally, a separate transformer is used on the entire RGB image as a whole (without splitting color channels) to capture image-wide context, preserving the original correlations between the color channels. Furthermore, to the best of current knowledge, this work is the first to enforce consistency between the enhanced images and their ground truth counterparts by passing them through the image encoder of a pretrained CLIP [140] model and minimizing the L2 loss between them.

3.2.1 Proposed Model

Let an underwater image dataset having M paired images $\mathcal{D} = \{(I_d^j, I_{gt}^j)\}_{j=1}^{M}$, where I_d and I_{gt} denote the degraded and ground-truth images respectively. Then, given an image I_d which is degraded due to several factors like color distortion, haziness, and low contrast, the goal is to obtain a corresponding enhanced image I_e using proposed novel model $I_e = f(I_d)$,

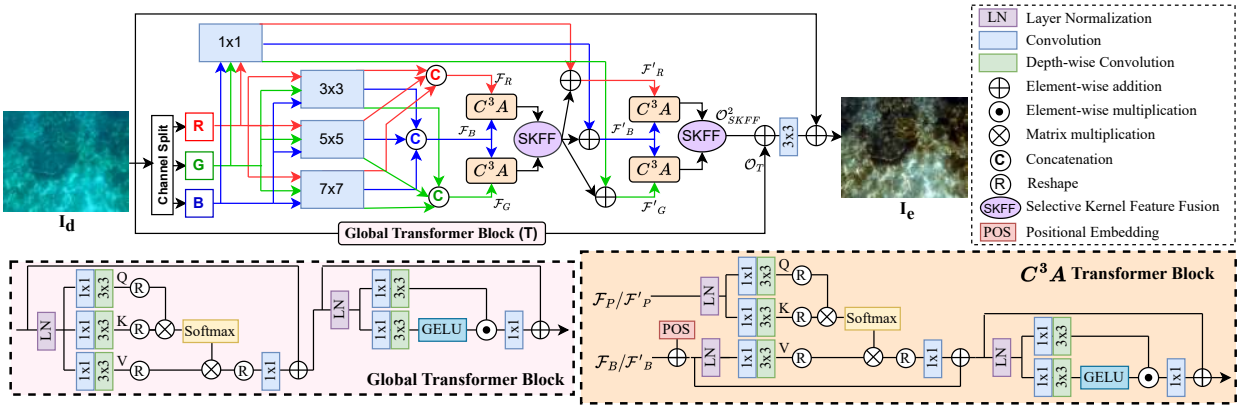


Figure 3.15: *Architecture of X-CAUNET. The model accepts a degraded underwater image as input and produces a corresponding visually-enhanced one. R-G-B denote the three image channels, with color-coded arrows indicating the channel-specific paths. P = R or G channel.*

reducing the aforementioned noise.

3.2.1.1 Cross-color channel attention transformer (C^3A)

The non-uniform attenuation of light in underwater scenarios can be attributed to wavelength, due to which different color components are scattered at different rates as water depth increases. Previous works have undertaken the CNN approach of handling the degraded images, largely with receptive fields of equal sizes. Sharma et al. [72] acknowledged that the blue color traverses the longest underwater due to its relatively shorter wavelength and proposed to use progressively bigger receptive fields in CNNs for the R-G-B channels. However, there exists no explicit associativity between the visual features of the three channels, undermining region-wise correlations between color channels. This work proposes a message-passing mechanism between the R-G-B channels to obtain correlated feature representations. First, the input I_d is splited into R-G-B channels (I_R, I_G, I_B). Then, local context is extracted from each channel using convolutional layers having three different sizes of receptive fields - $3 \times 3, 5 \times 5$, and 7×7. Consequently, a set of visual features $\{\mathcal{V}_{iR}, \mathcal{V}_{iG}, \mathcal{V}_{iB}\}_{i=3,5,7}$ are obtained, where $\mathcal{V}_{iC}$ denotes the $i \times i$ convolutional output for color channel C. Color-wise features for C are obtained by concatenating the visual features from different receptive fields, i.e., $\mathcal{F}_C = \mathcal{V}_{3C} \bigodot \mathcal{V}_{5C} \bigodot \mathcal{V}_{7C}$.

X-CAUNET proposes two cross-color channel attention (C^3A) transformers having the

same architecture for message passing between color-wise features. As evident from Figure 3.15, each C^3A performs a cross-attention between two color-channel features. Since blue color attenuates the slowest, it should retain more global information relative to others, and hence, this work keeps $\mathcal{F}_B$ as a common input to both C^3A_{BR} and C^3A_{BG}. Considering the former transformer, the operations performed are as follows (same for the latter as well). First, a positional embeddings ρ is added to the blue-channel features:

$$\Omega_B = \mathcal{F}_B + \rho \tag{3.16}$$

The message tokens for blue and red channels are obtained and passed through a cross-attention module:

$$\Delta_{BR} = CrossAttn(LN(\Omega_B), LN(\mathcal{F}_R)) \tag{3.17}$$

A residual connection is added to the resultant features:

$$\Theta_{BR} = \Delta_{BR} + \Omega_B \tag{3.18}$$

and followed by a Gated-Dconv Feedforward Network (GDFFN) inspired by [141] (LN = Layer Normalization):

$$\Psi_{BR} = GDFFN(LN(\Theta_{BR})) + \Theta_{BR} \tag{3.19}$$

Instead of simply concatenating the outputs Ψ_{BR} and Ψ_{BG}, this work capitalizes on the SKFF module [142] to obtain a refined version of the fused features at a relatively lower computational cost. To extract higher-level attention, another series of C^3A and SKFF blocks are deployed but with different inputs this time. I_R, I_G, and I_B pass through a 1×1 convolution, and the number of output channels are same as the output channels from

$\mathcal{O}^1_{SKFF}$ so that they can be added element-wise:

$$\mathcal{F'}_C = \mathcal{V}_{1C} + \mathcal{O}^1_{SKFF}, \;\; C \in \{R, G, B\} \tag{3.20}$$

where $\mathcal{V}_{1C}$ denotes output of 1×1 convolution on color-channel C. With the blue-color features common again to the C^3A transformers, it extract high-level attention features (following Eqs. 3.16 to 3.19 with inputs $\mathcal{F'}_R, \mathcal{F'}_G, \mathcal{F'}_B$), fusing them via $SKFF_2$ to get $\mathcal{O}^2_{SKFF}$.

3.2.1.2 Aggregating global context

To complement the attention cues obtained by merging visual features from individual color channels, this work adds another transformer to extract global information considering the whole image without channel split. This is an attempt to retain inter-channel consistencies that might be lost by splitting the RGB channels. Unlike C^3A, this global transformer (T) follows a traditional architecture given by [141] (Figure 3.15). Its outputs is denoted as $\mathcal{O}_T$. The final enhancement features are expressed as a weighted combination $\mathcal{Z} = \alpha \cdot \mathcal{O}^2_{SKFF} + \beta \cdot \mathcal{O}_T$, where α and β are learnable parameters. Since this work should finally produce an output image of the same dimensions as the input, it passes $\mathcal{Z}$ through a 3×3 convolution to adjust the output channels and add it to the degraded image to yield the enhanced image, i.e., $I_e = CONV(\mathcal{Z}) + I_d$.

3.2.1.3 Loss functions

Commonly used loss functions like L2 loss induce blurry artefacts in the enhanced images, motivating this work to use an L1 loss function instead:

$$l_1 = \frac{1}{N} \sum_{j=1}^{N} \| I^j_e - I^j_{gt} \|_1 \tag{3.21}$$

where N is the total number of training images. A noteworthy aspect of image enhancement is that the enhanced image should preserve the structural similarity with the

corresponding degraded image, and this work employs the SSIM for this purpose:

$$l_{SSIM} = \frac{1}{N}\sum_{j=1}^{N}(1 - SSIM(I_e^j, I_{gt}^j)) \tag{3.22}$$

Inspired by the recent success of large-scale visual-language models, X-CAUNET exploits the pretrained image encoder of CLIP [140] to extract powerful visual embeddings and proposes to minimize a new loss function:

$$l_{CLIP} = \frac{1}{N}\sum_{j=1}^{N}\|CLIP(I_e^j) - CLIP(I_{gt}^j)\|_2^2 \tag{3.23}$$

The overall objective function hence becomes:

$$\mathcal{L} = \lambda_1 \cdot l_1 + \lambda_2 \cdot l_{SSIM} + \lambda_3 \cdot l_{CLIP} \tag{3.24}$$

with experimentally set values of $\lambda_1 = 1, \lambda_2 = 0.5$, and $\lambda_3 = 0.02$.

Table 3.13: *UIQM comparison on U45 dataset. The first, second, and third best performances are represented in red, blue, and green, respectively.*

Method	FE	UDCP	FGAN	RB	RED	IBLA	WSCT	CycleGAN	AGCycleGAN	Proposed
UIQM	2.984	2.339	3.158	3.101	2.979	2.401	2.890	3.138	3.183	3.287

Table 3.14: *Comparison with the state-of-the-art on three datasets across six different evaluation metrics.*

Methods	UIEB					SUIM-E					UIEB Challenge	
	PSNR	SSIM	MS-SSIM	LPIPS	UIQM	PSNR	SSIM	MS-SSIM	LPIPS	UIQM	BRISQUE	UIQM
UDCP [127]	13.026	0.545	0.769	0.283	1.922	12.074	0.513	0.742	0.270	1.648	29.658	1.566
IBLA [50]	19.316	0.690	0.855	0.233	2.108	18.024	0.685	0.849	0.209	1.826	24.972	2.142
CBF [57]	20.771	0.836	0.890	0.189	3.318	20.395	0.834	0.884	0.194	3.003	29.213	2.810
UGAN [31]	23.322	0.815	0.932	0.199	3.432	24.704	0.826	0.941	0.190	2.894	25.118	2.662
FUnIE-GAN [32]	21.043	0.785	0.890	0.173	3.250	23.590	0.825	0.913	0.189	2.918	24.743	2.768
SGUIE-Net [75]	23.496	0.853	0.926	0.136	3.004	25.987	0.857	0.945	0.153	2.637	27.320	2.527
DWNet [72]	23.165	0.843	0.929	0.162	2.897	24.850	0.861	0.940	0.133	2.707	31.160	2.269
Proposed	24.121	0.871	0.939	0.135	3.132	24.721	0.886	0.947	0.121	2.855	23.980	2.712

3.2.2 Experiments

Proposed model is trained with samples from the UIEB [30] and SUIM-E [75] datasets. Besides, it tests the proposed model on the UIEB challenge set [30] and U45 [125] dataset to demonstrate the model generalization performance further. All experiments are performed using the PyTorch framework on a single Nvidia A100 GPU with a batch size of 5. The objective function is optimized with an Adam optimizer with a learning rate of 2e-4.

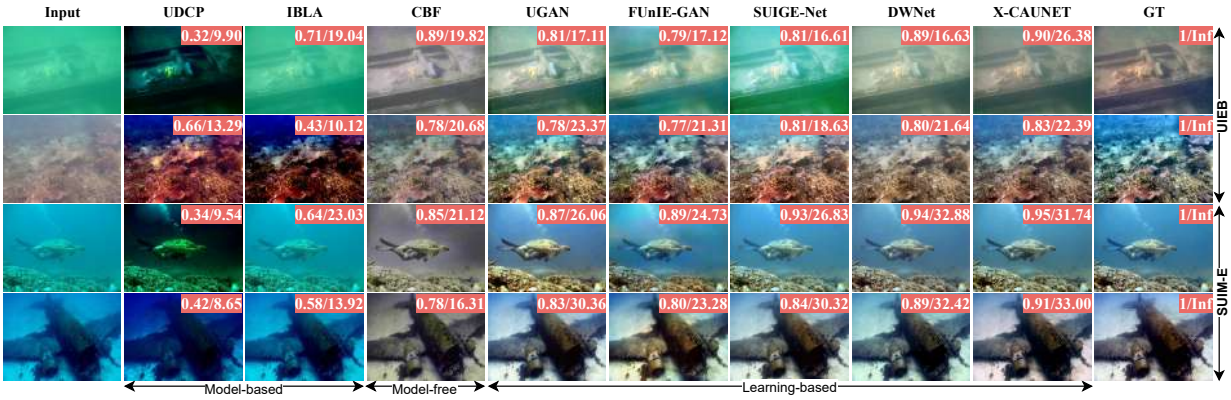


Figure 3.16: *Visual comparison with existing methods, with SSIM/PSNR values attained shown inside the red boxes.*

The quantitative comparison on UIEB, SUIM-E, and UIEB-challenge shown in Table 3.14 demonstrates that X-CAUNET substantially outperforms the previous SOTA. Specifically, X-CAUNET achieves 2.66%, 2.11%, and 0.75% relative gains on PSNR, SSIM, and MS-SSIM in the UIEB dataset. While on LPIPS, the proposed model has a lesser value, indicating a good perceptual quality of the enhanced image. On the UIEB challenge test set, X-CAUNET obtains SOTA with BRISQUE and comparable results with UIQM metrics. Table 3.13 shows the superiority of X-CAUNET on the U45 dataset. Figure 6.4 illustrates that after enhancement, X-CAUNET effectively removes haziness and corrects the greenish or bluish tone (also evident from Figure 3.17a).

Several ablation studies have been undertaken considering two aspects. With respect to model components, this work tries: **CS1:** removing the transformer block (T); **CS2:** choosing $\mathcal{F}_G$ as common input in C^3A instead of $\mathcal{F}_B$; **CS3:** choosing $\mathcal{F}_R$ as common input in C^3A instead of $\mathcal{F}_B$; **CS4:** one-to-one channel-kernel mapping, passing red, green and blue channels only via $3 \times 3, 5 \times 5$, and 7×7 convolutional kernels respectively; **CS5:** without

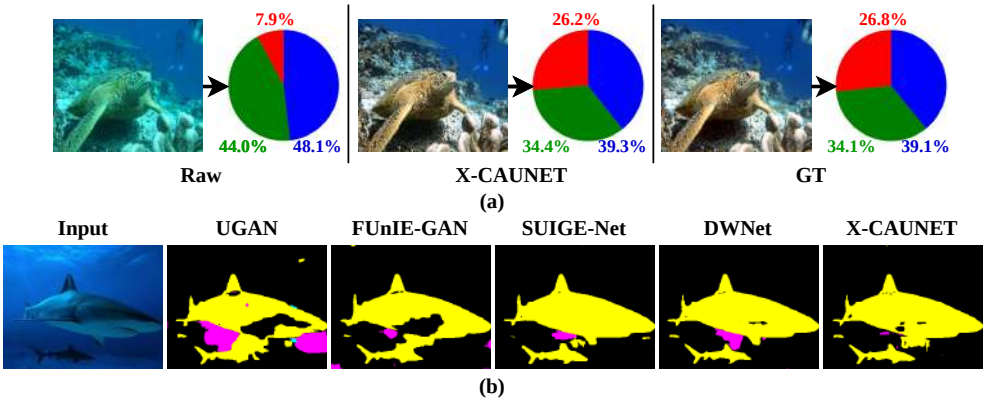


Figure 3.17: *(a) Color-channel dependencies in raw underwater images; (b) Impact on semantic segmentation by using the enhanced image produced by X-CAUNET as compared to previous UIE methods.*

5×5 and 7×7 convolutional blocks; **CS6**: without the 7×7 convolutional block; and **FM:** Full Model. Considering the loss terms, X-CAUNET is trained with: **LS1:** L1 loss only; **LS2:** L1 and CLIP losses; and **LS3:** L1, CLIP and SSIM losses. Figure 7.6 justifies the component utilities (corresponding visual comparisons are shown in Figure 3.19) and the collective usefulness of different losses.

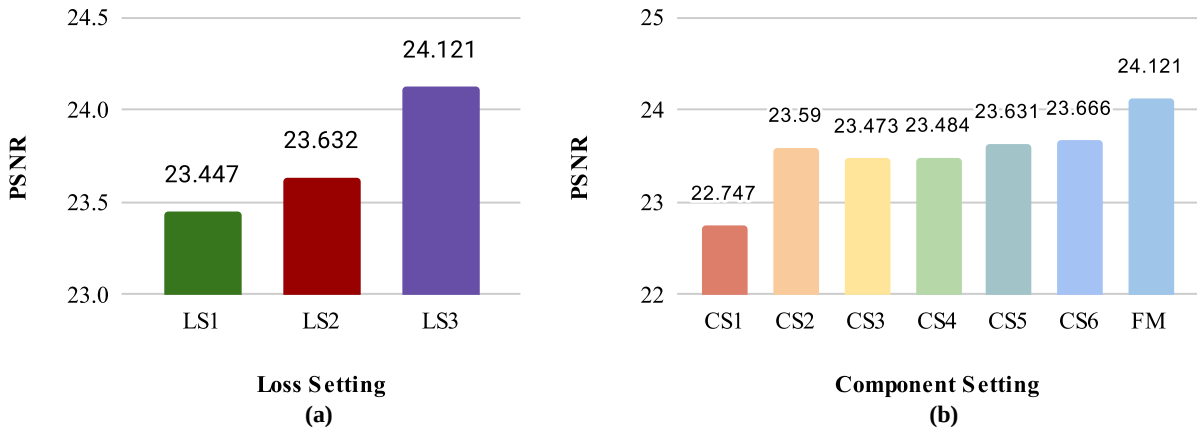


Figure 3.18: *Ablation studies with different Loss Settings (LS) and different Component Settings (CS) of the proposed model.*

3.3 Summary

In this chapter, first, a novel CNN-based light-weight image enhancement model, called Lit-Net, is proposed specifically for underwater images. In this model, there are two main

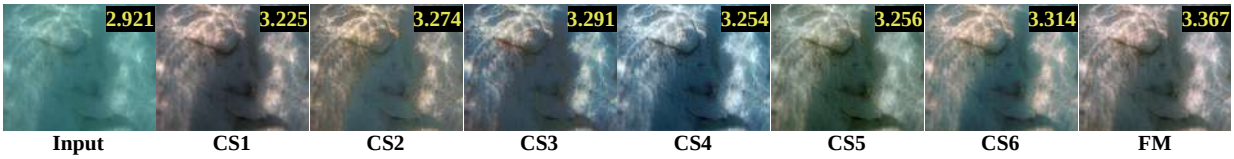


Figure 3.19: *Visual results of the proposed model with different component settings from the ablation study along with respective UIQM values obtained.*

modules, namely the MRAN and the MSAN. Different receptive fields (i.e., different convolution kernels) have been used to achieve multi-resolution image analysis while retaining the original input resolution. In the MSAN module, spatial attention-based skip connections are used between respective encoder-decoder layers to get both semantically and spatially rich feature representations. This encoder design facilitates the learning of diverse features while reducing the model's complexity. Furthermore, a weighted color channel-specific L1 loss is used to enhance the performance of Lit-Net. Experiments on UIE show that the proposed Lit-Net outperforms the state-of-the-art methods. Ablation studies are also conducted to validate the contributions of each module. It has also been experimentally demonstrated that images enhanced with the proposed model achieve significant improvements in the performance of various advanced visual tasks, such as underwater object detection and semantic segmentation.

Building upon the previous CNN-based model, the second work proposed X-CAUNET, a transformer-based framework that leverages the local context of individual color channels in underwater images while incorporating global attention cues from the entire image through learnable weights. The cross-attention mechanism enables effective message passing between the contextual features of different color channels, explicitly recognizing their correlations. Additionally, it is demonstrated that global attention from the whole image enhances the channel-specific features, resulting in improved image enhancement. The outcomes of this approach are more visually appealing compared to previous works.

Although the proposed models achieve promising results in underwater image enhancement, relying solely on the RGB color space limits color accuracy and stability. Underwater images often suffer from complex color distortions that are difficult to correct using RGB

information alone. In the next chapter, additional color spaces are incorporated to better separate color and brightness information, resulting in clearer images with more consistent and natural-looking colors. Furthermore, the combination of spatial and frequency domain information enables underwater image enhancement to preserve fine details while effectively correcting global degradations such as color cast and haze. This integrated approach improves texture clarity, edge sharpness, and overall visual quality.

The first work (Lit-Net) and the second work (X-CAUNET) presented in this chapter have been published in The Visual Computer journal and the ICASSP 2024 conference, respectively.

"The best way to predict your future is to create it."

~Abraham Lincoln

4

Dual Feature Transformations for Ocean-water Image Enhancement

Building on the progress of color channel-specific feature learning, the second contributory chapter focuses on enriching feature representations by leveraging dual color spaces and both spatial and frequency domain information. The first work utilizes dual-color space interactions, and the second work uses dual-domain feature modeling, enabling more robust enhancement.

4.1 UEnhancer: Spatially Enriched Transformer with Dual-color Information for Underwater Image Enhancement

Recently, the remarkable progress of deep learning across various computer vision domains has brought innovative strategies and viewpoints to the field of UIE [143]. Frequently, deep learning-based techniques leverage networks initially devised for other visual tasks [144,145] and adapt them for enhancing underwater images. Even though significant progress has been made, the UIE task still has scope for improvement. The primary reason is that existing deep UIE models overlook the inherent domain-specific knowledge in underwater imaging models.

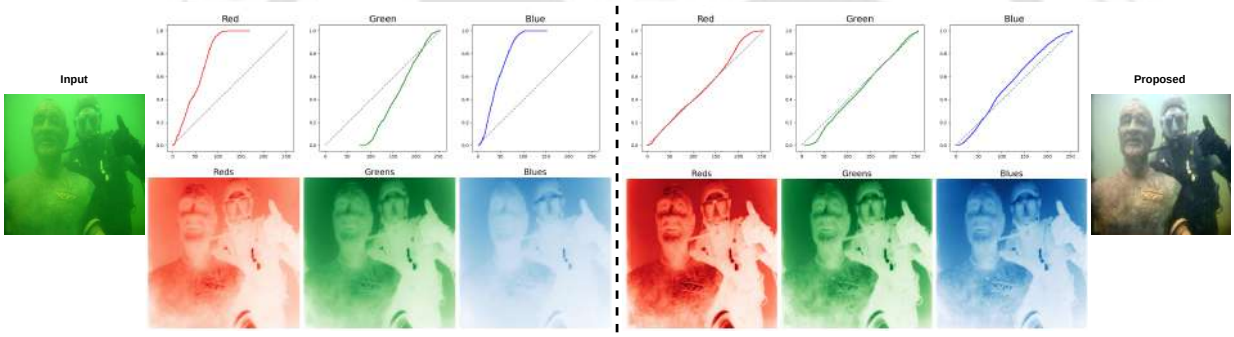


Figure 4.1: *R-G-B channel-wise image visualization and cumulative distribution of degraded and enhanced (by proposed method) image.*

This work aims to tackle the challenges of color cast problems, blurred details, low contrast, and detail loss in underwater images. This is achieved by employing sophisticated encoder features and capitalizing on the benefits of transformer-based methods. As previous studies [141, 146] have pointed out, the transformer has a limitation in grasping local dependencies. To address this, a novel spatial attention enhancer module is introduced into the transformer to capture local context. In contrast to previous deep models [72, 75, 83] that only depend on features extracted from the RGB color space, this approach involves exploring the feature representation using a dual-color (RGB and HSV) space network. By adopting this approach, this work significantly enhances the generalization ability of

deep networks while seamlessly integrating the unique attributes of both color spaces into a cohesive framework. Such an approach leads to improved performance and more robust handling of color casts and low-contrast issues in underwater images. Furthermore, a multi-scale frequency-based SSIM loss function is designed to preserve both the local structure and contrast in high-frequency regions.

Figure 4.1 demonstrates a general idea of what the individual color channels and their cumulative distribution functions (CDFs) look like. It can be seen that all three channels of degraded (input) images are quite far from the idealized straight line. The proposed model can substantially enhance the quality of the underwater image, mitigating the problems of color casts, blurred details, and low contrast, and resulting in visually enhanced outputs. The main contributions of this work can be outlined as follows:

1. This work proposes a novel model, UEnhancer, designed specifically for underwater image enhancement. The model utilizes a dual-color transformer-based hierarchical encoder-decoder architecture to handle different underwater image challenges, such as color cast problems and low contrast issues.

2. A spatial attention enhancer (SAE) is introduced within the transformer block to grasp the local context, which is crucial for the UIE task.

3. A dual-color interaction block is proposed as a message-passing mechanism between the RGB-HSV color space to obtain correlated feature representations.

4. A multi-scale frequency-based SSIM loss is proposed to preserve the local structure, the contrast in high-frequency regions, and detailed information.

4.1.1 Proposed Method

This section first presents the pipeline of a dual-color-based transformer network for the UIE task and then discusses in detail its key components.

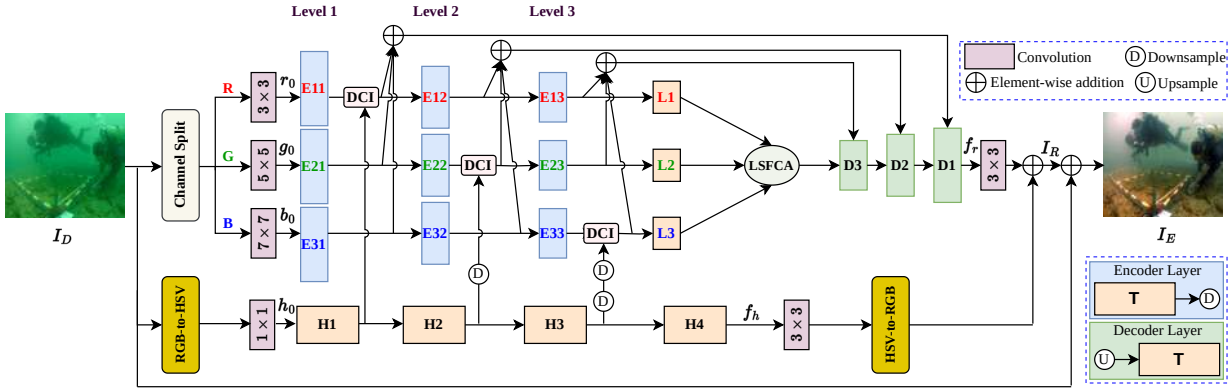


Figure 4.2: *Overview of the proposed UEnhancer, which utilizes the dual color space (RGB and HSV) and learns enriched local-global contextual feature representation for UIE.*

4.1.1.1 Overall Pipeline

The overall pipeline of UEnhancer is shown in Figure 4.2. Given a degraded underwater image $I_D \in \mathbb{R}^{3\times H\times W}$, it first splits the image into R, G, and B channels and parallelly transforms the RGB image into HSV color space. UEnhancer applies a 3×3, 5×5, 7×7, and 1×1 convolutional layer into R, G, B and HSV to obtain low-level feature embeddings r_0, g_0, b_0, and $h_0 \in \mathbb{R}^{C\times H\times W}$. Then the feature maps r_0, g_0, and b_0 are passed through 3-level three different encoders. In each level, each encoder contains a transformer block (T) (refer to Figure 4.3) followed by a down-sampling operation (D) using the pixel-unshuffle. Feature map h_0 undergoes through a series of four transformer blocks, yielding the deep features of the HSV color space denoted as f_h. During the encoding process, it uses a message-passing mechanism between the R-G-B and HSV color spaces to obtain correlated feature representations. Next, the latent representation of three encoders is fused using the proposed latent space fusion with convolution attention (LSFCA), and a 3-level single decoder is used to reconstruct the encoded feature and get f_r. Each decoder contains an upsampling layer (pixel-shuffle) and a transformer block similar to the encoder. To facilitate the reconstruction procedure, the encoder features are merged with the decoder features through a skip connection. Next, a convolutional layer is employed to refine features f_r and f_h, convert the HSV to RGB, and generate the residual image I_R. Lastly, the degraded underwater image is combined with the residual image to get the enhanced image

$$I_E = I_D + I_R.$$

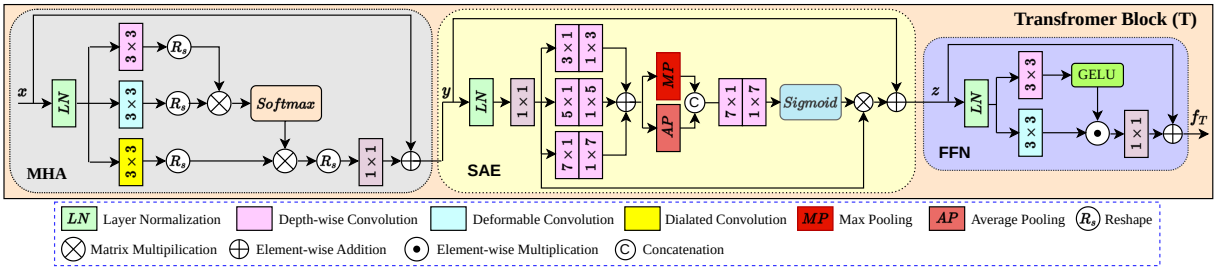


Figure 4.3: *The components of the transformer block. MHA, SAE, and FFN indicate the multi-head attention, spatial attention enhancer, and feed-forward network, respectively.*

4.1.1.2 Transformer Block

The standard transformer block architecture computes the self-attention (SA), and it is computationally costly. The computational and memory demands of the key-query dot product interaction increase quadratically with the input's spatial resolution, i.e., $O(W^2H^2)$. Hence, applying SA to image enhancement tasks, which frequently entail high-resolution images, becomes ineffective due to the computations. To ease this problem, [141] reduces the complexity to a linear scale by proposing Multi-Dconv Head Transposed Attention (MDTA). Moreover, capturing local dependencies is crucial for image enhancement tasks, as the pixels surrounding a degraded pixel can provide valuable contextual information for restoring its clean version. However, the transformer fails to effectively capture these local dependencies. To tackle this problem, this work modifies the traditional transformer block as illustrated in Figure 4.3, which captures both long-range dependencies (global context) and valuable local (spatial) context. A Spatial Attention Enhancer block is built between the multi-head attention and feed-forward network.

$$f_T = FFN(SAE(MHA(x))) \tag{4.1}$$

Multi-head attention computes cross-covariance across channels, effectively creating an attention map that implicitly captures global contextual information. In MHA, this work introduces deformable and dilated convolutions along with depth-wise convolution before

calculating feature covariance to generate a global attention map. Deformable convolution allows for the adaptive adjustment of receptive fields, which will be beneficial for enhancing fine details in images. It will capture and emphasize local structures and textures, leading to improved perceptual quality. It helps in avoiding over-smoothing and retaining important features, where the depth-wise convolution is used to globally enhance the entire image without a strong emphasis on local details. From a layer normalized tensor (LN), MHA first generates query (Q), key (K), and value (V) projections by applying a 3×3 depth-wise (f_{dp}), 3×3 deformable (f_{df}), and 3×3 dilated (f_{da}) convolution with k dilation factor (k = 1,2, and 3 for R, G, and B channel, respectively). Underwater images commonly exhibit greenish and bluish tints, so this work increases the value of k on these channels to capture more global information. Let x be an intermediate feature, then

$$Q = f_{dp}(LN(x)),\ K = f_{df}(LN(x)),\ V = f_{da}(LN(x)) \tag{4.2}$$

Next, the query (Q_*) and key (K_*) projections are reshaped and perform the dot product to generate an attention map. To achieve multi-head SA, the channels are partitioned into multiple heads, allowing each head to learn attention maps independently and in parallel. The MHA's output is denoted as y, where

$$y = V_* \otimes SAttn(Q_*, K_*) + x \tag{4.3}$$

Spatial attention enhancer captures the local context by leveraging the inter-spatial relationship of features. By attending to regions with important information, it can selectively enhance details while avoiding over-amplification of noise and achieve a more vibrant and visually appealing color balance in the enhanced image. In SAE, it first applies 1×1 convolution to a layer-normalized tensor, followed by three parallel depth-wise strip convolutions. This work selects depth-wise strip convolutions because they are lightweight and effectively emulate standard 2D convolutions with receptive-field sizes of 3×3, 5×5, and 7×7, for this only need a pair of 3×1 and 1×3 convolutions, 5×1 and 1×5 convolutions, and 7×1 and 1×7 convolutions. Next, all the strip features are added, perform both max and

average pooling operations along the channel axis, and concatenate the results to produce an adequate feature descriptor. Finally, a pair of 7×1 and 1×7 convolutions is used to produce the spatially attentive feature map. Let y be the feature map after the MHA, then the SAE can be computed as follows:

$$z = SAE(y) \tag{4.4}$$

FFN is designed as the element-wise product of two parallel paths; one path features a 3×3 depth-wise convolution followed by GELU, and the other path employs a 3×3 deformable convolution. Similar to MHA, here also one depth-wise convolution is replaced with a deformable convolution. Let z be the feature map obtained from the SAE, then

$$f_T = FFN(z) \tag{4.5}$$

4.1.1.3 Dual-color Interaction (DCI) block

In contrast to terrestrial scene images, underwater images exhibit a broader spectrum of color deviations, ranging from bluish or greenish tones to yellowish hues. The wide range of colors poses a significant challenge for conventional network architectures. To operate various color spaces in the enhancement algorithm, a dual-color interaction block is proposed, where the same images have diverse visual representations across multiple color spaces. Figure 4.4(a) shows the basic idea of the proposed dual-color interaction, where the fusion involves the feature of HSV and the feature of R-G-B in different scales. While the R-G-B components in RGB color space exhibit a high correlation, causing it to be sensitive to variations induced by luminance alterations, occlusion, and shadows, the HSV color space offers a valuable inherent depiction, disentangling hue, saturation, brightness, and contrast attributes with greater clarity and robustness. To amalgamate the distinctive properties of both color spaces for UIE, this work integrates the unique characteristics of RGB and HSV

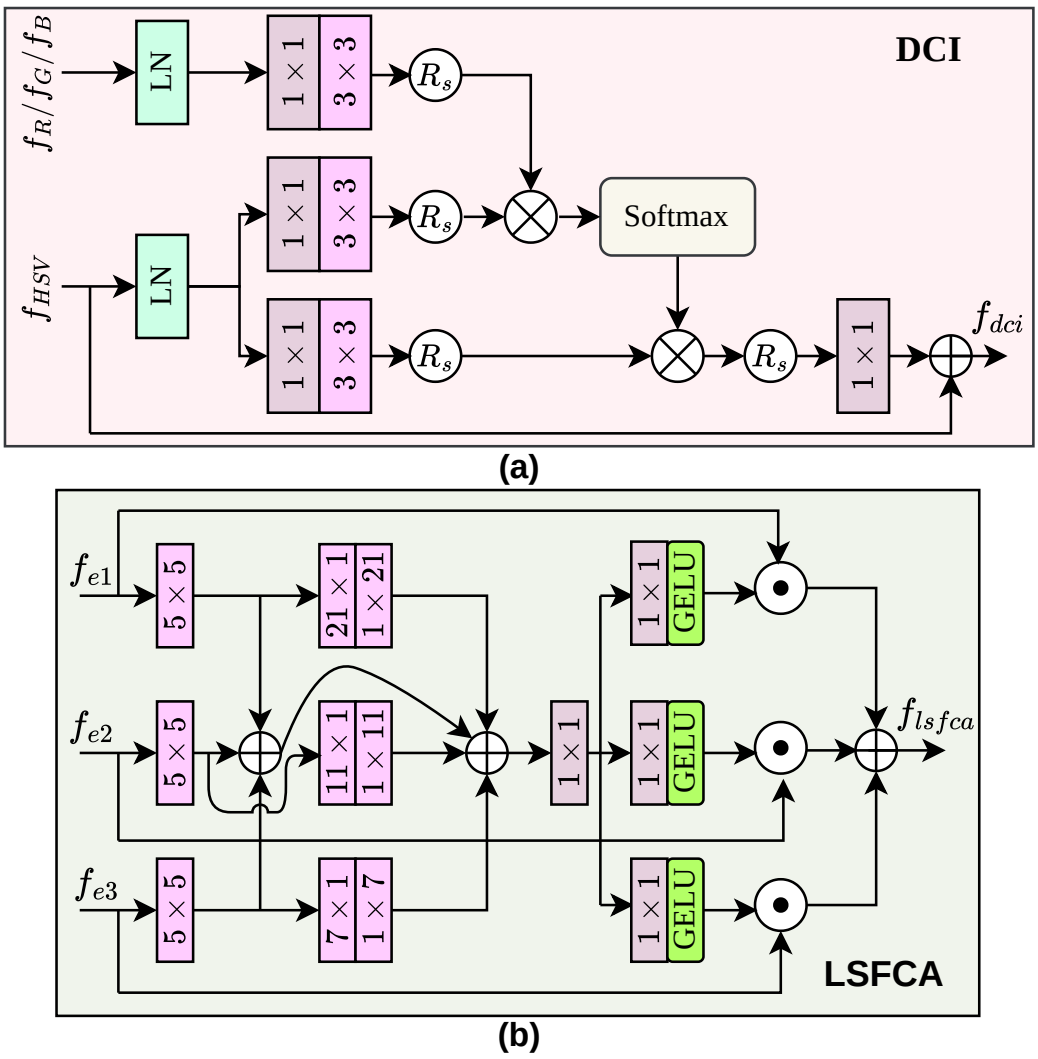


Figure 4.4: *Illustration of proposed (a) Dual-color interaction (DCI) and (b) Latent space fusion with convolution attention (LSFCA) block.*

color spaces within a cohesive deep framework. Furthermore, the color disparity between two points with a slight spatial separation in RGB color space may exhibit significant variation when viewed in HSV color space. Hence, leveraging several color space embeddings enhances the precision of measuring color deviations in underwater images, giving a broader perspective on the variations across different color representations.

This work proposes three interaction blocks with the same architecture for message passing between different color-wise features. As evident from Figure 4.4(a), each interaction

block performs a cross-attention between two color features. Since the blue color attenuates the slowest, it should retain more global information relative to others. Hence, it performs the cross-attention between the B color channel and HSV color space at the end of the 3rd level encoder (refer to Figure 4.2). In the case of red color, it captures the local context and performs the cross-attention between R and HSV at the 1st level encoder. The interaction block can be expressed as follows:

$$f_{dci} = DCI(f_{cc}, f_{HSV}), \quad cc = R, G, B \tag{4.6}$$

4.1.1.4 Latent Space Fusion with Convolution Attention

Latent space fusion, combining features extracted from diverse encoder layers, is an omnipresent detail of current network architecture. The fusion technique is commonly implemented through simple summation or concatenation; the effectiveness of these approaches may vary and may not always produce optimal results. Linear techniques such as addition and concatenation inherently lack contextual information. To mitigate this issue, it employs various kernel sizes to aggregate context using a lightweight attention mechanism. Instead of employing a cross-attention mechanism, this work has used a progressive multi-scale convolutional attention consisting of three integral parts: a depth-wise convolution for accruing local context, a multi-path depth-wise strip convolution to grasp the contextual information in multiple scales, and a 1×1 convolution to model the interrelationships within various channels intricately. The proposed LSFCA block is depicted in Figure 4.4(b). Let f_{e1}, f_{e2}, and f_{e3} be the three latent features of three different encoders, and then the output of fused latent space (f_{LSFCA}):

$$f_{LSFCA} = LSFCA(a, b, c) \tag{4.7}$$

4.1.1.5 Loss Function

UEnhancer uses two loss functions: L2 loss (L_2) and multi-scale frequency-based SSIM loss (L_{MSF}). It employs a linear combination of the aforementioned loss functions to train

UEnhancer, with the total loss defined as follows:

$$L_T = L_2 + \lambda_1 \times L_{MFS} \tag{4.8}$$

where λ_1 is empirically set to 0.5.

L2 loss. L2 loss is used to preserve image detail, as noise and outliers can adversely affect the performance of the UIE model. This approach enhances model optimization and addresses image color distortion issues. The L2 loss is formulated as follows:

$$L_2 = \frac{1}{N}\sum_{i=1}^{N} \|I_E^i - I_{GT}^i\|_2^2 \tag{4.9}$$

where, I_{GT} and I_E indicates the ground-truth and predicted image, respectively.

Multi-scale frequency cube-based SSIM (L_{MFS}) loss. Next, it uses L_{MFS} loss to maintain the structural similarity and preserve the contrast in high-frequency regions. This work has used k-level decomposition of the image using Haar wavelet transform to calculate the L_{MFS} loss. The multi-level nature of the transform allows for the separation of signal and noise at different scales, which improves the overall image quality. Algorithm 1 shows the L_{MFS} loss between the predicted enhanced image and ground-truth image. For a single batch, it can be computed as:

$$L_{MFS} = \frac{1}{N}\sum_{i=1}^{N} 1 - MSFL(I_E^i, I_{GT}^i, k) \tag{4.10}$$

where MSFL(.) refers to the loss's algorithm.

4.1.2 Experiments

4.1.2.1 Experimental Settings

Datasets. This work conducts experimentations on multiple datasets to demonstrate the effectiveness of UEnhancer. It has used EUVP dataset [32], LSUI dataset [74], SUIM-E dataset [75], UIEB dataset [30] for training and testing. Apart from that, UIEB challenge

Algorithm 1 MFS loss's algorithm

procedure MSFL(I_E, I_{GT}, k) ▷ I_E, I_{GT}, and k refer to the enhanced image, ground-truth image, and number of DWT levels, respectively.
  $LL^0 \leftarrow I_E$
  $LL_g^0 \leftarrow I_{GT}$
  $S \leftarrow 0$
  for $i \leftarrow 1$ `to` k **do**
    $LL^i, LH^i, HL^i, HH^i \leftarrow 2dDWT(LL^{i-1})$
    $LL_g^i, LH_g^i, HL_g^i, HH_g^i \leftarrow 2dDWT(LL_g^{i-1})$
    $S_{LH} \leftarrow SSIM(LH^i, LH_g^i)$
    $S_{HL} \leftarrow SSIM(HL^i, HL_g^i)$
    $S_{HH} \leftarrow SSIM(HH^i, HH_g^i)$
    $S_i \leftarrow \frac{1}{3}(S_{LH} + S_{HL} + S_{HH})$
    $S \leftarrow S + S_i$
  end for
  return $S \leftarrow S/k$ ▷ The L_{MFS} loss value.
end procedure

set [30], U45 dataset [125], and UCCS dataset [147] are used for testing.

Evaluation Metric. To ensure a rigorous and comprehensive comparison between the proposed UEnhancer and existing SOTA methods [33, 72, 74, 79], this work evaluates their performance using well-established with and without reference-based image quality metrics. These include MSE, SSIM, LPIPS, and PSNR, as well as MS-SSIM, and BRISQUE. Additionally, it incorporates domain-specific metrics tailored for underwater imagery, such as the UIQM and UISM. To indicate a good measurement, the models desire lower MSE, LPIPS, and BRISQUE values and higher values for other metrics. This meticulous evaluation ensures a comprehensive understanding of the model's performance across various aspects of

Table 4.1: *Quantitative evaluation on E515 testset with the SOTA methods. Best, second best, and third best results are highlighted in red, blue, and green colors, respectively.*

Method	LPIPS↓	MSE↓	PSNR	SSIM	MS-SSIM	BRISQUE↓	UIQM	UISM
UGAN [31]	0.220	0.503	26.551	0.807	0.936	35.859	2.896	6.833
P-UGAN [31]	0.223	0.503	26.549	0.805	0.935	35.099	2.93	6.816
F-GAN [32]	0.212	0.544	26.220	0.792	0.924	30.912	2.971	6.892
F-GAN-UP [32]	0.246	0.875	25.224	0.788	0.925	34.070	2.935	6.853
D-SESR [109]	0.206	0.515	27.081	0.803	0.940	35.179	3.099	7.051
WaveNet [72]	0.173	0.449	28.654	0.835	0.950	30.856	3.042	7.051
Ushape [74]	0.187	0.370	26.822	0.811	0.949	35.648	3.052	6.843
Unformer [148]	0.216	0.561	27.858	0.824	0.944	36.952	2.997	6.791
Spectroformer [77]	0.198	0.614	27.225	0.822	0.943	33.901	2.998	6.852
CDF_UIE [149]	0.216	0.541	28.196	0.831	0.946	33.647	3.029	6.854
GuidedUIR [150]	0.202	0.553	28.010	0.836	0.945	35.636	3.033	6.824
Proposed	0.148	0.345	30.014	0.876	0.963	32.720	2.982	7.014

image quality assessment.

Training Process. The proposed UEnhancer model is executed using the open-source framework PyTorch. All experiments were executed on an Ubuntu 20.04 OS with a single Nvidia A100 GPU. It leveraged the Adam optimizer in the training phase, initializing the learning rate at 0.0002. A batch size of 4 was kept throughout the training process. Empirically, the k value is set to 3 in the k-level DWT (refer to Algorithm 1). Also, a dropout layer is used at the end of every transformer block to effectively reduce overfitting, allowing the model to generalize new data better. The dropout probability is set to 0.1.

SOTA Methods. In the context of enhancing the underwater images, the proposed UEnhancer is evaluated against various SOTA, including Fusion [68] ($CVPR-12$), U-DCP [127] ($ICCVW-13$), Retinex based [63] ($ICIP-14$), Gdehaze [49] ($ICASSP-16$), IBLA [50] ($TIP-17$), ULAP [56] ($PCM-18$), CBF [57] ($TIP-18$), RGHS [59] ($MMM-18$), UGAN [31] ($ICRA-18$), WaterNet [30] ($TIP-19$), UIE-DAL [20] ($CVPRW-19$), F-GAN [32] ($RAL-20$), DeepSESR [109] ($RSS-20$), Ucolor [33] $TIP-21$, SGUIE-Net [75] ($TIP-22$), WaveNet [72] ($TOMM-23$), Ushape [74] ($TIP-23$), WFI2-net [79] ($CVPR-24$), UIE-Convformer [151] ($TETCI-24$), Unformer [148] ($TAI-24$), Spectroformer [77] ($WACV-24$), CDF_UIE [149] ($TGRS-25$), and GuidedUIR [150] ($CSVT-25$).

The proposed technique uses the same training and test sets as other techniques to ensure a fair comparison. The resolution of the training and testing images was consistent across all methods. Additionally, to maintain uniformity in the evaluation process, it used an identical environment setting (Python/Matlab) to compute the evaluation metrics.

4.1.2.2 Comparisons with SOTA schemes

Quantitative Evaluation. First, it performs the quantitative comparison on E515, U90, S110, and L400. Table 4.1 reports the quantitative comparison of the EUVP dataset (E515). It is observed from the table that the proposed method outperforms GuidedUIR [150] with a 60.28% lower MSE score and 7.15% and 4.78% higher PSNR and SSIM scores, demonstrating UEnhancer's superior capability to enhance underwater images. This finding suggests

Table 4.2: *Quantitative evaluation on U90 test set with the SOTA works for UIE.*

Method	LPIPS↓	PSNR	SSIM	MS-SSIM	BRISQUE↓	UIQM
U-DCP [127]	0.283	13.026	0.545	0.769	24.134	1.922
Gdehaze [49]	0.309	15.378	0.671	0.777	23.929	2.520
IBLA [50]	0.233	19.316	0.690	0.855	23.711	2.108
ULAP [56]	0.256	19.863	0.724	0.865	25.113	2.328
CBF [57]	0.189	20.771	0.836	0.890	20.535	3.318
UGAN [31]	0.199	23.322	0.815	0.932	27.011	3.432
P-UGAN [31]	0.192	23.550	0.814	0.933	25.382	3.396
F-GAN [32]	0.173	21.043	0.785	0.890	18.522	3.250
SGUIE-Net [75]	0.136	23.496	0.853	0.926	24.607	3.004
WaveNet [72]	0.162	23.165	0.843	0.929	24.863	2.897
U-shape [74]	0.220	21.084	0.744	0.895	24.128	3.161
WFI2-net [79]	0.124	23.860	0.873	-	-	-
UIE-Convformer [151]	-	22.130	0.880	-	-	-
Unformer [148]	0.158	23.628	0.861	0.943	23.404	3.146
Spectroformer [77]	0.108	24.482	0.884	0.949	23.667	3.096
CDF_UIE [149]	0.138	24.413	0.870	0.945	20.669	3.158
GuidedUIR [150]	0.147	23.412	0.873	0.944	21.484	3.239
Proposed	0.104	25.279	0.893	0.952	23.551	3.092

that UEnhancer can adaptively recover brightness and intricate details, significantly improving color and spatial regions. Table 5.1 represents the results of the UIEB dataset (U90). Compared with the previous best result, UEnhancer improves SSIM, PSNR, and MS-SSIM scores 1.01%, 3.25%, and 0.31%, respectively. The LPIPS score is 3.84% less than the Spectroformer [77] result, indicating high perceptual similarity and good image quality. It correlates well with human judgments of image similarity. Apart from that, this work also gets comparable UIQM and BRISQUE scores. Next, the SUIM dataset, featuring diverse scenarios, further improves the adaptability of the proposed method. From Table 4.4, one can see that the proposed method achieved the best score on four out of six evaluation metrics on the SUIM-E dataset (S110). The LSUI dataset includes a wide variety of underwater environments. Similarly, in the case of the LSUI test set (L400), UEnhancer increases 19.97% PSNR and 2.15% SSIM over Ushape, as shown in Table 5.2. Also, it can be seen that from Table 4.6 it has fewer parameters than most of the existing methods except GuidedUIR, with the highest results.

Next, this work conducted tests on UC60, U45, and UCCS. The outcomes are presented

Table 4.3: *UIQM comparison on C60, U45, and UCCS dataset with the SOTA works.*

	U-DCP [127]	IBLA [50]	ULAP [56]	F-GAN [32]	Ucolor [33]	SGUIE-Net [75]	WaveNet [72]	Ushape [74]	TUDA [152]	WFI2-net [79]	Proposed
C60	1.304	1.935	1.815	2.506	2.481	2.527	2.269	2.783	2.415	-	2.729
U45	2.049	1.710	2.283	2.563	3.148	3.112	3.128	2.923	-	3.181	3.187
UCCS	2.691	2.555	2.208	3.074	2.983	-	3.089	2.874	3.175	-	3.200

Table 4.4: *Quantitative evaluation on S110 test set with the SOTA works for UIE.*

Method	LPIPS↓	PSNR	SSIM	MS-SSIM	BRISQUE↓	UIQM
U-DCP [127]	0.270	12.074	0.513	0.742	22.788	1.6482
Gdehaze [49]	0.355	14.339	0.599	0.743	20.175	2.255
IBLA [50]	0.209	18.024	0.685	0.849	20.957	1.826
ULAP [56]	0.231	19.148	0.744	0.871	21.250	2.115
CBF [57]	0.194	20.395	0.834	0.884	21.115	3.003
UGAN [31]	0.190	24.704	0.826	0.941	20.288	2.894
P-UGAN [31]	0.188	25.050	0.827	0.943	18.768	2.901
F-GAN [32]	0.189	23.590	0.825	0.913	22.560	2.918
SGUIE-Net [75]	0.153	25.987	0.857	0.945	25.927	2.637
WaveNet [72]	0.133	24.850	0.861	0.941	20.757	2.707
Ushape [74]	0.213	22.647	0.783	0.917	19.602	2.873
Unformer [148]	0.143	25.097	0.879	0.954	20.892	2.877
Spectroformer [77]	0.102	25.473	0.889	0.958	19.782	2.775
CDF_UIE [149]	0.131	25.592	0.883	0.954	19.952	2.872
GuidedUIR [150]	0.114	25.796	0.892	0.953	19.632	2.887
Proposed	0.100	27.500	0.902	0.963	19.551	2.857

in Table 4.3. As no reference images exist for these test sets, the non-reference evaluation metric UIQM is used. A higher value of UIQM demonstrates that the proposed method is more compatible with human visual perception. From Table 4.3, it can be observed that UEnhancer achieved 0.92% relative gain on UIQM on the U45 dataset.

Qualitative Evaluation. This work shows the visual comparison in Figure 4.5, 4.6, 4.7, 4.8, and 4.9. Comparing the results, it can be noticed a) the proposed method adeptly recovers complex details and enhances contrast in foreground and background without sacrificing overexposed or underexposed parts of the image, and b) additionally, it reveals vivid and natural colors, rendering the enhanced results more realistic and visually appealing. In contrast to the C60 dataset, the results (refer to Figure 4.9) of U-DCP, ULAP, and SGUIE-Net fail to enhance the image's contrast. WaveNet generates the dark region in the

enhanced image. Ucolor and Ushape produce better results but exhibit limited effectiveness in scenarios where the contrast is extremely low. In the case of the U45 dataset, SGUIE-Net, ULAP, and U-DCP still exhibit the greenish tone after the enhancement. For a more comprehensive comparison of the enhancement of accurate reflection, this work shows the pixel value statistics in RGB color space of the enhanced results, as depicted in Figure 4.10. The pixel distribution clearly demonstrates that UEnhancer achieves a more balanced and favorable distribution than the other methods. Further, as demonstrated in Figure 4.12, the existing methods yield higher UIQM values but produce visually compromised images with a noticeable reddish tint and inadequate enhancement, which indicates that UIQM may not accurately assess image quality in some instances. Nonetheless, the proposed method produces images with bright reflections and natural effects without introducing undesirable artifacts.

The visualization in Fig. 4.11 shows the proposed framework design, which uses different kernel sizes for each color channel: red (R = 3), green (G = 5), and blue (B = 7). This approach addresses wavelength-dependent attenuation in underwater imaging. The encoder feature maps indicate that each channel captures distinct structures: red focuses on fine details, green preserves mid-scale features, and blue highlights broad haze patterns. This separation enhances feature extraction. In the decoder stages, scene structures and colors are restored, resulting in outputs that retain fine details and accurate colors. The activations suggest the network learns wavelength-specific representations, improving robustness against underwater degradation.

4.1.2.3 Ablation Study

This section discusses the impact of different components on the overall outcomes of the proposed UEnhancer.

Impact of loss functions. As illustrated in Figure 4.13, the UEnhancer without L_{MSF} loss produces fade results, which suggests that the L_{MSF} loss helps the generated image to contain identical characteristics as the ground-truth image, resulting in enhanced images

Table 4.5: *Quantitative evaluation on L400 test set with the SOTA works for UIE.*

Method	PSNR	SSIM	LPIPS↓
Fusion [68]	17.48	0.79	0.31
U-DCP [127]	11.89	0.59	0.22
Retinex based [63]	13.89	0.74	-
IBLA [50]	13.54	0.71	0.26
RGHS [59]	14.21	0.78	0.21
WaterNet [30]	17.73	0.82	0.24
UGAN [31]	19.79	0.78	0.19
F-GAN [32]	19.37	0.84	-
UIE-DAL [20]	17.45	0.79	0.28
Ucolor [33]	22.91	0.89	0.23
U-shape [74]	26.13	0.93	0.22
WFI2-net [79]	27.26	0.94	0.11
UIE-Convformer [151]	26.57	0.91	-
Unformer [148]	28.88	0.91	0.17
CDF_UIE [149]	28.25	0.89	0.16
GuidedUIR [150]	29.82	0.90	0.14
Proposed	31.35	0.95	0.10

Table 4.6: *Parameter comparison with SOTAs.*

	WaterNet	UGAN	Ucolor	Ushape	UIE-DAL	UIE-Convformer	Unformer	CDF_UIE	GuidedUIR	Proposed
Params (M)	24.81	57.17	157.4	65.6	18.82	25.9	7.31	15.90	1.14	4.86

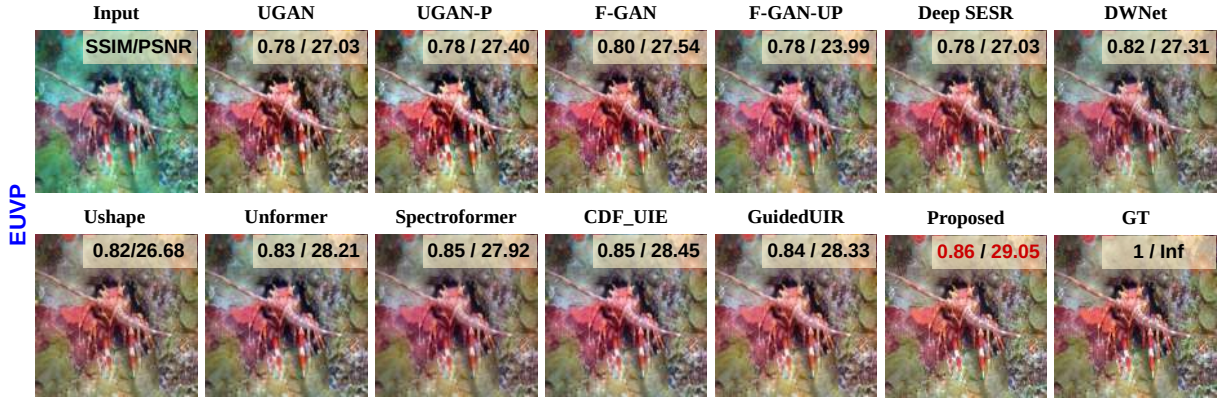


Figure 4.5: *Qualitative comparison with SOTAs on EUVP dataset.*

that appear realistic and useful in noise suppression and smoothing. Table 4.7 further shows that the experimental findings align consistently with Figure 4.13, as demonstrated via quantitative evaluation. The UEnhancer produces optimal results across all evaluation metrics under the combined supervision of the L_2 and L_{MSF} loss functions. This ensures

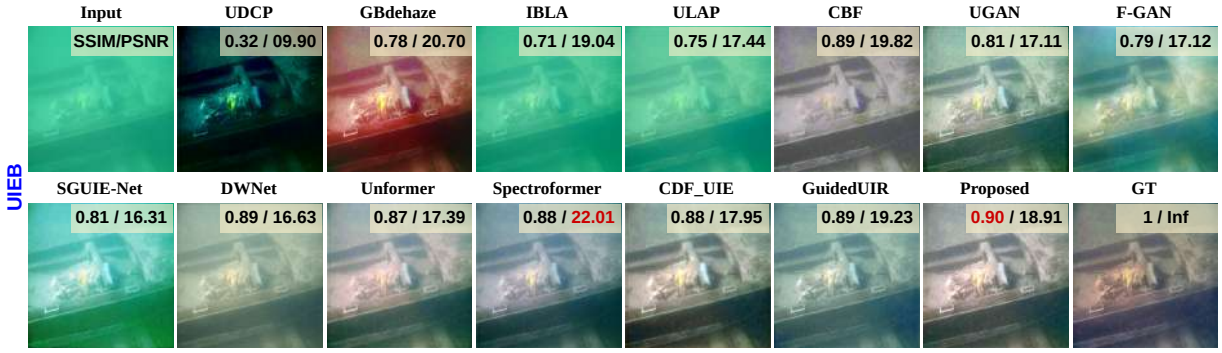


Figure 4.6: *Qualitative comparison with SOTAs on UIEB dataset.*

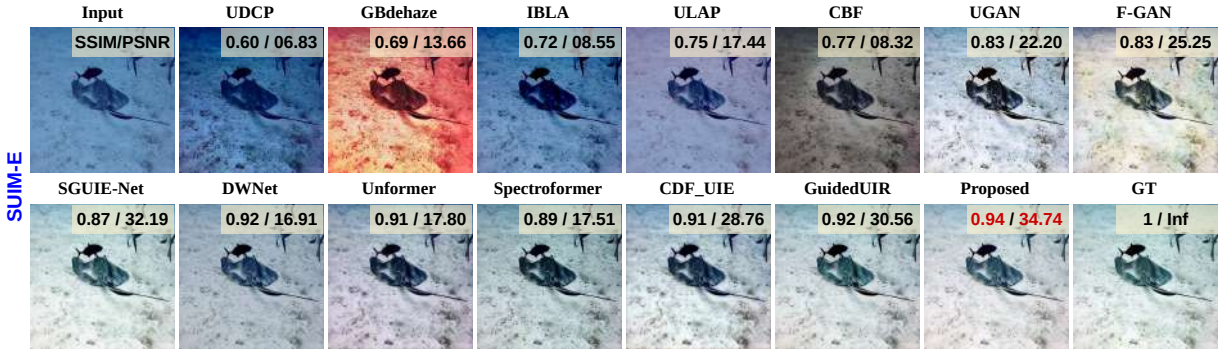


Figure 4.7: *Qualitative comparison with SOTAs on SUIM-E dataset.*

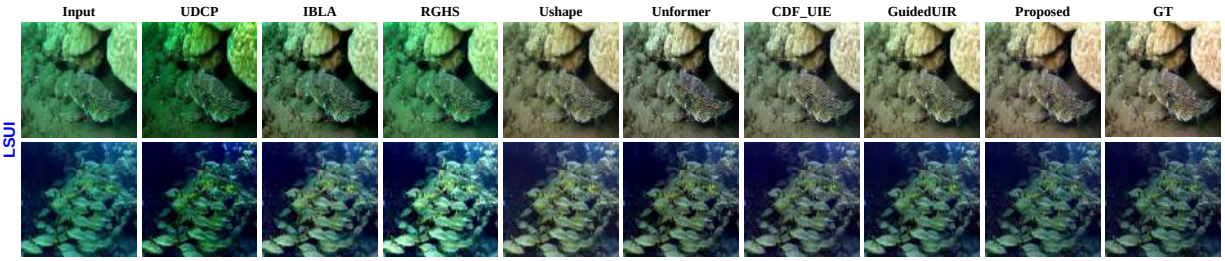


Figure 4.8: *Qualitative comparison with SOTAs on LSUI.*

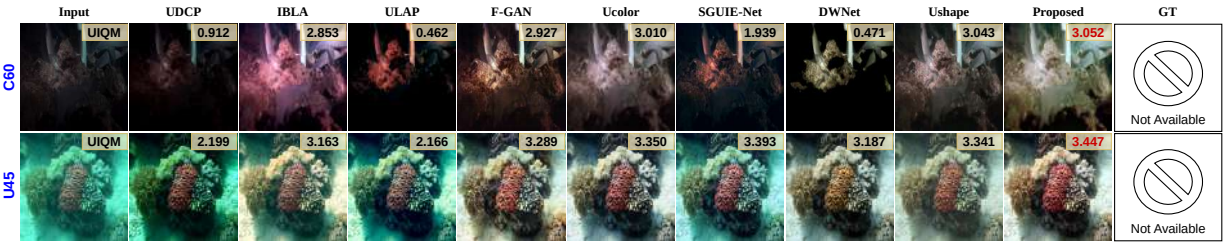


Figure 4.9: *Qualitative comparison with SOTAs on C60 and U45 datasets.*

that these specific loss functions are crucial for creating high-quality underwater images with improved contrast, color, and reduced noise.

Impact of Spatial Attention Enhancer. To elucidate the effectiveness of SAE,

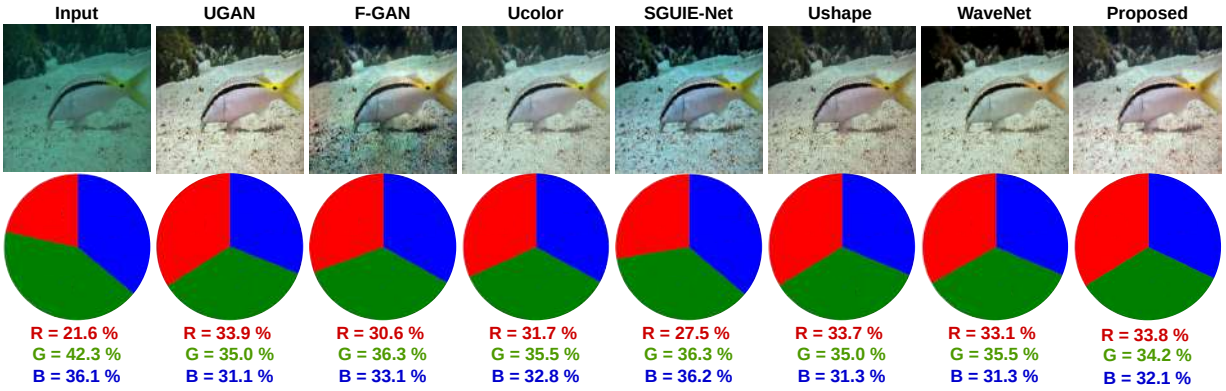


Figure 4.10: *Comparison of pixel distribution statistics of different methods through a pie chart.*

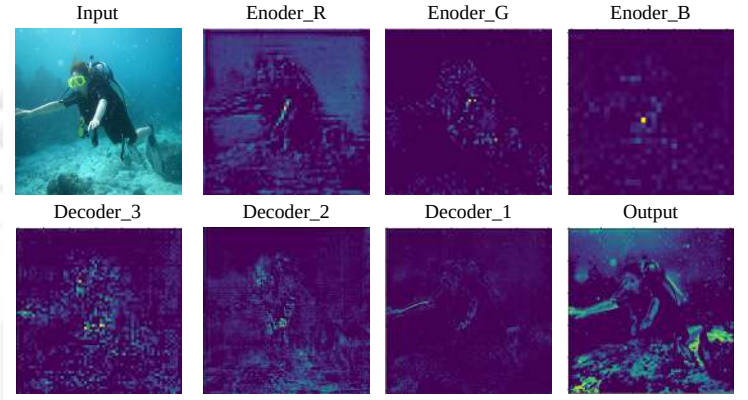


Figure 4.11: *Intermediate feature maps, showing channel-specific focus on fine edges (red), textures (green), and haze patterns (blue) that are progressively fused in the decoder for accurate enhancement.*

Table 4.7: *Effectiveness of various cost functions on U90.*

	l_2	$l_2 + \lambda_1\, l_{MFS}$
PSNR	24.478	25.279
SSIM	0.876	0.893
LPIPS	0.130	0.104
UIQM	3.280	3.092

UEnhacer is trained without an SAE block. As shown in Table 4.8, without SAE, it exhibits decreased performance in underwater image quality metrics. Additionally, the visual result of UEnhancer is better with SAE, which suggests the importance of the SAE block (refer to Figure 4.14).

Effect of dual-color space. The ablated model without HSV is demonstrated in Table 4.8 and Figure 4.14. While RGB color space is essential for representing colors in terms of

Figure 4.12: *Visual comparisons between proposed model and the existing model. The middle and right columns show the results of the existing and proposed methods, respectively.*

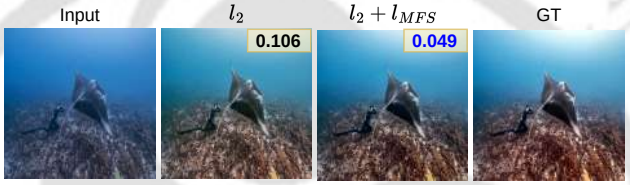


Figure 4.13: *Visual effect of two losses. The top corner indicates the LPIPS score.*

Table 4.8: *Effectiveness of various settings in the proposed UIE model on U90.*

Modules	PSNR	SSIM	LPIPS	UIQM
w/o SAE	24.56	0.882	0.121	3.053
MHA w/o f_{dc} and f_{da}	25.04	0.874	0.115	3.071
w/o HSV	24.47	0.872	0.125	3.043
w/o R-G-B encoder	24.63	0.885	0.111	3.068
w/o LSFCA (addition)	24.56	0.871	0.119	3.077
w/o LSFCA (concatenation)	24.53	0.877	0.115	3.056
Full model	25.279	0.893	0.104	3.092

their red, green, and blue components, HSV color space offers additional advantages, such as intuitive color adjustments, robustness to illumination changes, and efficient color segmentation, making it a valuable complement for underwater image enhancement. In contrast, a well-designed dual-color space embedding facilitates us in learning robust representations to enrich enhancement results.

Effect of R-G-B channel-wise encoder. The quantitative result in Table 4.8 shows that splitting the R-G-B color channel increases the performance in terms of PSNR and SSIM values. Also, it can be noticed from Figure 4.14 that the enhanced image is better

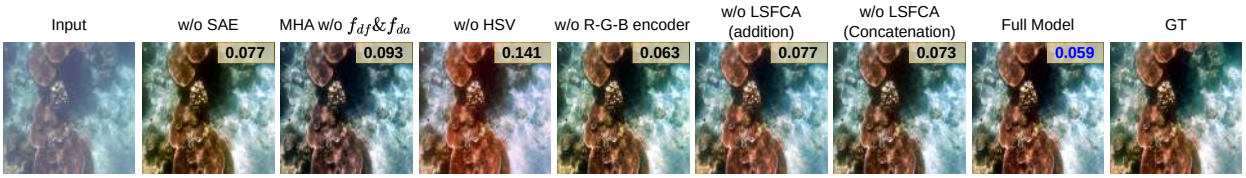


Figure 4.14: *Influence of various components in UEnhancer. The top corner indicates the LPIPS score.*

than that by training with a single RGB encoder. Thus, it is necessary to split the color channel and process it through the model to improve the quality of the final results.

Impact of LSFCA Block. To study the impact of LSFCA, this work constructs two ablation modules, "addition" and "concatenation". Table 4.8 presents their comparison on the UIEB dataset. It is evident that when compared to the linear process, the non-linear fusion process incorporating an attention mechanism consistently yields superior results. The result strongly demonstrates that LSFCA improves the visual (refer to Figure 4.14) and quantitative quality of underwater image enhancement.

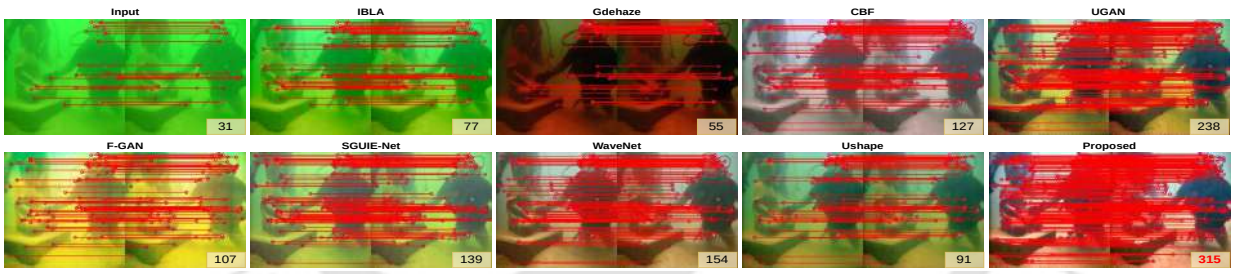


Figure 4.15: *The display of keypoint matching. Note that the proposed method achieves the highest number of keypoint matches.*

4.1.2.4 Impact on Various Applications

To validate the potential practical applicability of the proposed model, this work employs UEnhancer as a prepossessing technique for underwater image **keypoint matching**, **edge detection**, and **semantic segmentation** tasks, as shown in Figure 4.15, 4.17, and 4.16, respectively. Initially, it employs scale-invariant feature transform (SIFT) to compute the keypoints. It can be observed that the keypoint matching significantly increases after enhancing the images by proposed method. Subsequently, it employs Canny edge detection

to identify edges in the degraded and enhanced images. Compared to the degraded image and enhanced image by SOTA models, enhanced image by proposed work exhibits a greater number of localized edge features. Finally, it employs the SUIMNet [153] model to get the semantic segmented image. The semantic segmentation of the image enhanced by proposed model has increased the completeness of the structure compared to the others.

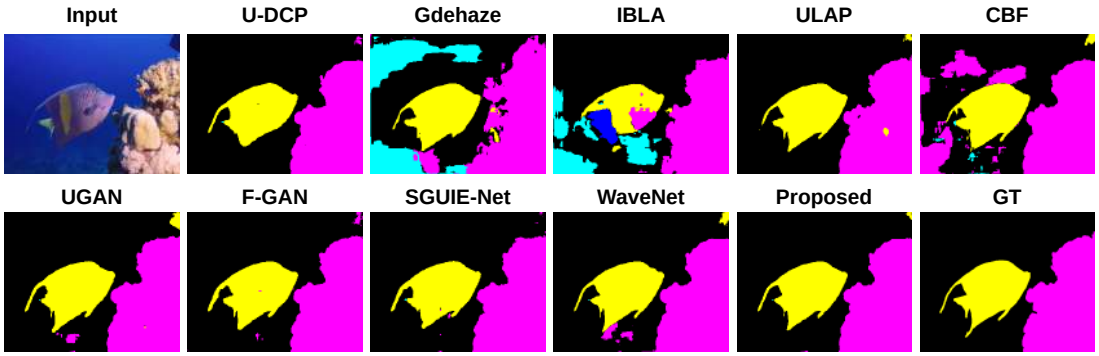


Figure 4.16: *Visual evaluation of semantic segmentation on SUIM dataset.*

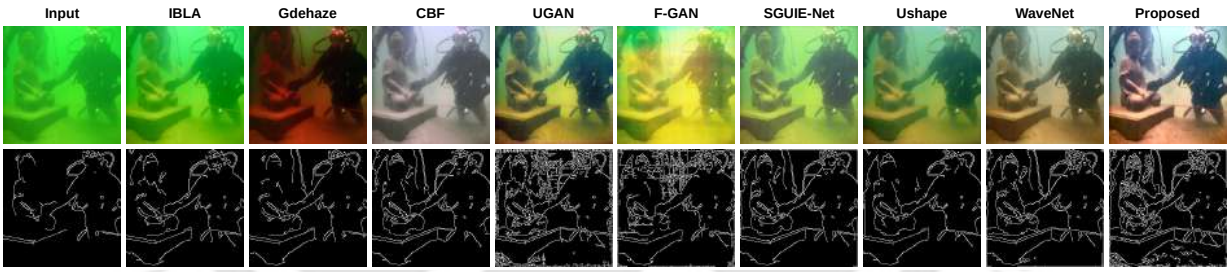


Figure 4.17: *Canny edge detection comparison between existing methods and proposed work.*

4.1.2.5 Generalization performance of the proposed method

To further assess UEnhancer model's performance on low-level visual tasks, this work focused on enhancing low-light images. It used the LOL dataset, which included 500 paired images of a low-light and a corresponding normal exposure version. Of these, 485 pairs were for the training, while the remaining 15 were for the test. The quantitative result on the LOL dataset is presented in Table 4.9. The UEnhancer model exhibits a pronounced superiority over the existing method, achieving a substantial performance enhancement across all evaluated metrics on the LOL dataset. Figure 4.18 illustrates that UEnhancer delivers impressive visual results, effectively balancing brightness enhancement with artifact minimization.

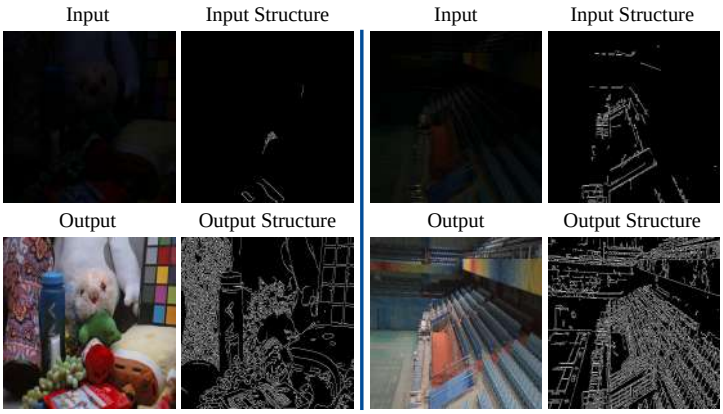


Figure 4.18: *Top row represents the input and corresponding structure of the input. The last row represents the output and corresponding structure of output by the proposed model.*

Table 4.9: *PSNR and SSIM comparison on LOL dataset with existing methods.*

	PSNR	SSIM
HWMNet [154]	24.55	0.857
Decoupled [155]	21.84	0.82
CLE-Diffusion [156]	25.51	0.825
LightingNet [157]	22.74	0.842
LL-UNet++ [158]	23.00	0.868
DSFPNet [159]	23.90	0.859
RIRO [160]	23.01	0.897
OptiRet-Net [161]	21.87	0.80
Proposed	25.82	0.901

4.2 D2Mamba: Dual Domain Guided Informed Search in State Space Model for Underwater Image Enhancement

Researchers have proposed various methods to improve enhancement, focusing on traditional and learning-based approaches. Traditional physics-based UIE methods [127, 162] attempt to reverse degradation using handcrafted priors but rely on manually tuned parameters and fail to generalize to complex underwater conditions. Learning-based approaches address this limitation by replacing manual priors with data-driven models. Early CNN-based UIE methods improved generalization but struggled to capture long-range dependencies, leading to limited performance. More recently, Transformer-based methods have shown

promise, as self-attention effectively models global dependencies and non-local similarities. However, they still face issue as global attention introduces quadratic computational cost. Most existing approaches focus primarily on the spatial domain, with limited exploration of frequency-domain information, which restricts the ability of deep models to fully exploit feature representations for high-quality UIE.

More recently, SSMs [94,163] and their recent variants, Mamba [97] and Mamba-2 [164], have emerged as powerful backbones for balancing global context modeling and computational efficiency. Their recursive formulation enables the capture of ultra-long dependencies, which are crucial for addressing complex degradation in underwater images. Moreover, Mamba's parallel scan algorithm ensures efficient processing on modern hardware. Based on Mamba's strengths, a few studies have utilized the Mamba framework to improve the quality of degraded underwater images, yielding promising results [101, 104]. However, they often rely on fixed scanning mechanisms [165] such as raster, cross, diagonal, or bi-directional scan, and often fail to handle irregular and complex underwater degradations. Raster, row, and column scans use a linear traversal that is efficient but does not adapt to varying content, leading to loss of critical structural details. While bi-directional and cross scans capture features from multiple directions and offer improved variability, they still lack the necessary adaptability to address the irregular patterns found in degraded underwater imagery, limiting their effectiveness in accurately reconstructing structures. Since traditional scanning methods are rigid and degradation-agnostic, there is a need for a scanning mechanism that adapts to the input's physical degradation, allowing more effective and richer feature propagation for underwater image enhancement.

To address the above limitations, this work proposes D2Mamba, a lightweight dual-domain framework for underwater image enhancement. It jointly processes spatial and frequency-domain representations to capture local textures and global spectral characteristics, while its Mamba-based backbone efficiently models long-range dependencies. In Mamba, unlike conventional scanning, it designs a scanning mechanism guided by actual degradation, such as hazy regions and color-attenuated areas, instead of following a fixed

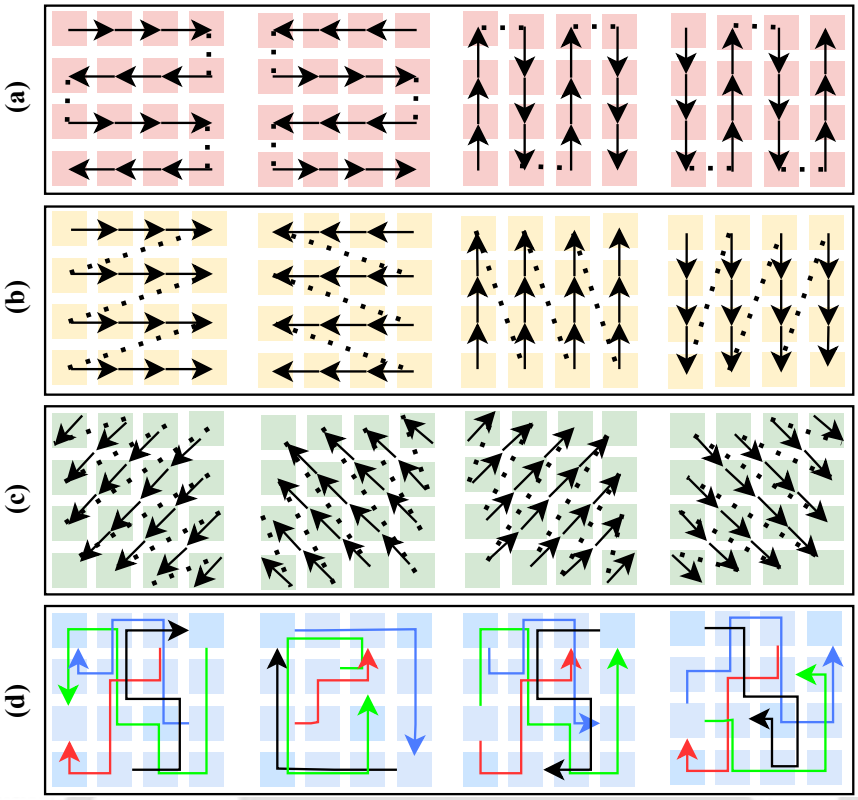


Figure 4.19: *Comparison of different scanning strategies for* MAMBA BLOCK. *Unlike (a) bidirectional, (b) raster, or (c) diagonal, proposed (d) A* guided GIFH traversal adaptively follows degradation-aware paths, enabling more effective and sparse feature propagation.*

order. This approach allows the traversal path to prioritize the more informative regions dynamically with relatively low overhead (refer to Figure 4.19). The scanning mechanism is designed to find the optimal path by modifying the A* algorithm with the Geodesic Information-Field Heuristic (GIFH). Additionally, D2Mamba uses a Spectral Wasserstein Attenuation Loss (SWAL) to align the color (spectral) distributions of restored images with natural image statistics. By computing the Wasserstein distance in Lab color space, SWAL enforces consistent global colors and perceptually realistic details, improving both color fidelity and visual quality in underwater image enhancement. The major contributions are:

- A lightweight Mamba-based framework is proposed that incorporates spatial and frequency domain information for efficient global context modeling and detail preservation.

- A physics-informed Geodesic Information-Field Heuristic (GIFH) is proposed that

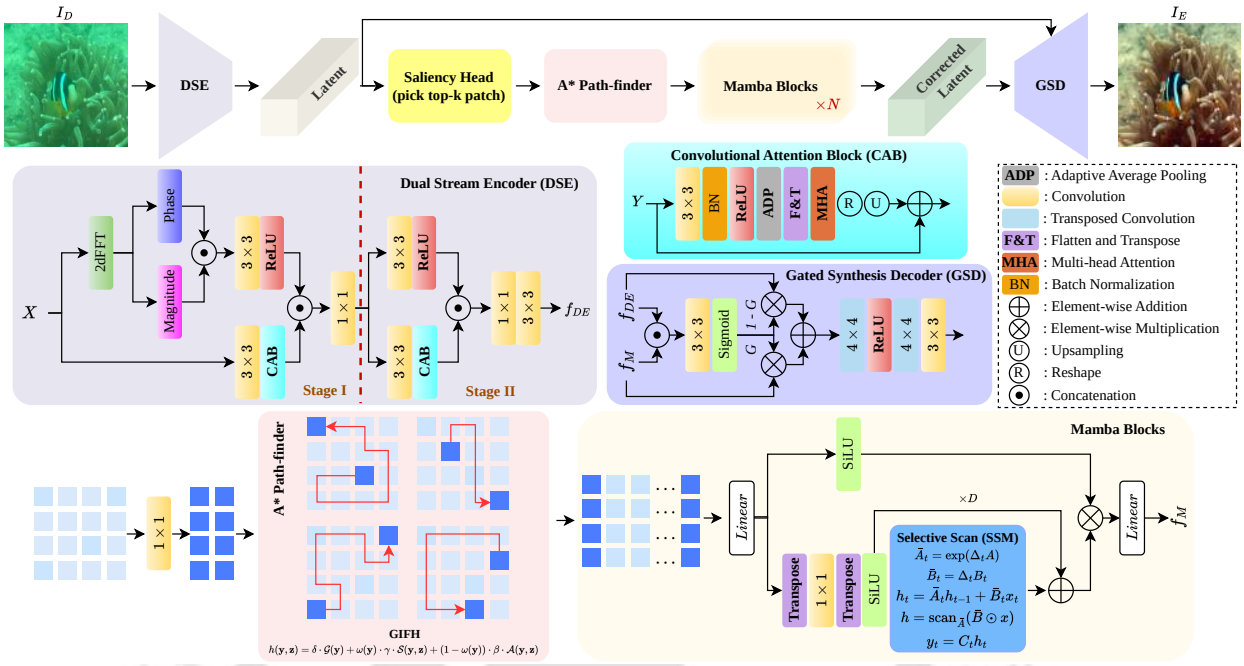


Figure 4.20: *Overall Architecture of D2Mamba. The network extracts spatial and frequency features via a dual-stream encoder, adaptively traverses them using A* guided GIFH, enhances sequential features with Mamba blocks, and reconstructs images through a gated decoder.*

adaptively guides feature traversal to handle spatially varying degradations.

- A Spectral Wasserstein Attenuation Loss (SWAL) is designed to align color distributions with natural statistics, ensuring perceptually consistent restoration.
- D2Mamba achieves state-of-the-art results on UIE benchmarks with significantly lower parameters and FLOPs than existing methods.

4.2.1 Proposed Method

The proposed D2Mamba network introduces a novel paradigm combining physics-informed pathfinding with selective state-space modeling. The architecture comprises four interconnected modules: Dual-stream encoder, Geodesic-guided pathfinding, Mamba-based sequential correction, and Gated synthesis decoder, as shown in Figure 4.20. The main idea is to enhance underwater images by finding the best paths through a learned set of features, where the local features represent degradation characteristics.

4.2.1.1 Overall Model Integration

Given a degraded underwater image $I_D \in \mathbb{R}^{3\times H\times W}$, this work first passes it through the Dual-Stream Encoder (DSE) to extract both local texture patterns and global spectral characteristics via progressive feature refinement and get the latent feature f_{DE}. The resulting latent feature map is then processed by a saliency-based node selector to identify top-k degraded regions:

$$\begin{aligned} \mathbf{S_h} &= Conv_{1\times1}(\mathbf{f}_{DE}) \in \mathbb{R}^{H/4\times W/4}, \\ \text{Nodes} &= topk(flatten(\mathbf{S_h}), k = 2P), \end{aligned} \tag{4.11}$$

where P denotes the number of enhancement paths in the Mamba block. The selected nodes are split into two regions according to saliency scores, and node pairs are formed by sampling one from each region, ensuring that paths connect distinct degraded areas. The A*-based pathfinder [166], guided by affinity costs and the geodesic information-field heuristic (GIFH), tracing a few adaptive paths and generating a sparse correction map where only path pixels receive strong context-aware updates. Path features are then zero-padded for uniform batch processing:

$$\mathbf{f}_{path} = ZeroPad([\mathbf{f}_{DE}[\mathbf{p}_{j,i}]]_{i=1}^{L_j}) \quad \in \mathbb{R}^{P\times L_{\max}\times C}, \tag{4.12}$$

where j, L_j, and $\mathbf{p}_{j,i}$ represent the path index, path length, and node index, respectively. These path features are processed by the Mamba block for selective correction, followed by spatial re-integration through overlap averaging:

$$f_M = \frac{1}{\text{count}[r,c]} \sum_{j:(r,c)\in\mathcal{P}_j} \mathbf{f}_{\text{corrected}}[j, \text{step}_{r,c}]. \tag{4.13}$$

Finally, the corrected latent representation f_M is fused with the original latent f_{DE} in a Gated Synthesis Decoder (GSD) to generate the output. The corrected latent contains a sparse representation that is mapped back to the global image space using transpose convolution layers. When the decoder identifies a confident correction along a restored edge,

it shares that information with nearby pixels, gradually refining the surrounding areas. This step-by-step process produces the enhanced underwater image $I_E \in \mathbb{R}^{3\times H\times W}$.

4.2.1.2 Dual-Stream Encoder (DSE)

Traditional image processing methods apply uniform techniques, overlooking the varying degradation levels found underwater. In contrast, the proposed dual-stream encoder processes spatial and frequency domains simultaneously with two stage architecture, capturing local textures and global characteristics while refining feature space. In stage I, DSE takes I_D as input and decomposes the spectrum into physically meaningful components by 2D FFT.

$$\mathbf{M}_{fft}, \mathbf{\Phi}_{fft} = \text{FFT2D}(I_D). \tag{4.14}$$

The magnitude component $\mathbf{M}_{fft}$ encodes wavelength-dependent attenuation effects, while the phase component $\mathbf{\Phi}_{fft}$ captures scattering-induced distortions. This work creates a compact 2-channel frequency representation by averaging (μ) across color channels:

$$\mathbf{I}_{freq} = [\mu_{ch}(\mathbf{M}_{fft}), \mu_{ch}(\mathbf{\Phi}_{fft})] \in \mathbb{R}^{B\times 2\times H\times W}. \tag{4.15}$$

Then, the frequency and spatial streams are processed in parallel to produce compatible feature representations. The spatial stream employs convolution-attention block (CAB) for capturing long-range spatial dependencies producing $s_1 \in \mathbb{R}^{B\times C\times H\times W}$, while the latter uses a 3×3 convolution followed by ReLU producing $f_1 \in \mathbb{R}^{B\times C\times H\times W}$, where C represents the feature dimension. CAB employs efficient attention computation with fixed downsampling to 8×8 spatial resolution, achieving computational efficiency, while preserving spatial relationships through residual connections and bilinear upsampling. Then, it combines spatial and frequency features through concatenation followed by 1×1 convolution while reducing dimensionality. This fused representation retains strengths of both domains, balancing texture detail with global frequency-aware context.

$$\begin{aligned} \mathbf{f}^1_{cat} &= [\mathbf{s}_1; \mathbf{f}_1] \in \mathbb{R}^{B\times 2C\times H\times W}, \\ \mathbf{f}^1_{fused} &= Conv_{1\times 1}(\mathbf{f}^1_{cat}) \in \mathbb{R}^{B\times C\times H\times W}. \end{aligned} \tag{4.16}$$

Stage II applies synchronized 2× downsampling along with channel expansion to f^1_{fused}, similar to the previous stage. This process creates a high-level feature representation with more compact spatial dimensions and enriched semantic encoding, resulting in $f_{DE} \in \mathbb{R}^{B\times 4C\times H/4\times W/4}$.

4.2.1.3 Enhanced A* Pathfinding with Affinity-Based Edge Costs

Classical A^* search [166] computes a path from node x to node z using the cost function $f(x,z) = g(x,y)+h(y,z)$, where g calculates the cost of reaching a neighbor and h estimates the remaining distance to the target. Building on this intuition, this work adapts A^* for degraded feature traversal in underwater images by determining edge costs directly in the feature space. The transition cost $g(x,y)$ between two nodes is calculated using feature affinity:

$$g(x,y) = 1 - CosSim(\mathbf{f}_{DE}[x], \mathbf{f}_{DE}[y]), \tag{4.17}$$

where cosine similarity ($CosSim$) between $L2$ normalized feature vectors intrinsically captures semantic similarity while ensuring non-negative costs required for A^* optimality. To enhance efficiency, all edge costs within an 8-connected neighborhood are precomputed in a vectorized manner. The costs are constrained within the range of $[0, 2]$, where higher costs indicate greater feature dissimilarity. This allows the search process to prioritize paths that traverse more coherent and less degraded features.

4.2.1.4 Geodesic Information-Field Heuristic (GIFH)

To guide A^* in finding optimal traversal paths, it introduces a novel Geodesic Information-Field Heuristic (GIFH) that takes into account important physical phenomena related to underwater image degradation. Rather than relying solely on geometric distances, GIFH incorporates three complementary factors: geodesic curvature, scattering variance, and ab-

sorption distance. This combination results in a physics-informed measure of path cost as shown:

$$\begin{aligned} h(\mathbf{y}, \mathbf{z}) = \delta \cdot \mathcal{G}(\mathbf{y}) + \omega(\mathbf{y}) \cdot \gamma \cdot \mathcal{S}(\mathbf{y}, \mathbf{z}) \\ + (1 - \omega(\mathbf{y})) \cdot \beta \cdot \mathcal{A}(\mathbf{y}, \mathbf{z}), \end{aligned} \tag{4.18}$$

where δ, γ, β are learnable parameters processed through softplus activation to ensure positivity.

Geodesic component ($\mathcal{G}$) [167] measures how suddenly the degradation changes around a location using Sobel gradients [168]. Large gradients indicate "sharp bends" in the degradation field, discouraging paths from crossing highly unstable regions.

$$\mathcal{G}(\mathbf{y}) = \frac{1}{C} \sum_{c=0}^{C-1} \|\nabla \mathbf{f}_{DE}[\mathbf{y}]\|_2. \tag{4.19}$$

The scattering component ($\mathcal{S}$) evaluates the differences in high-frequency variances among nodes [169]. This promotes the movement of paths through areas with similar textural variability as the objective, highlighting how scattering has a significant impact on high-frequency details.

$$\mathcal{S}(\mathbf{y}, \mathbf{z}) = \mathrm{Var}_{HF}(\mathbf{z}) - \mathrm{Var}_{HF}(\mathbf{y}). \tag{4.20}$$

Local variance is computed using the standard formula $\mathbb{E}[\mathbf{f}_{HF}^2] - (\mathbb{E}[\mathbf{f}_{HF}])^2$ (where, the high-frequency $\mathbf{f}_{HF} = \mathbf{f}_{DE}[:, :, C/2 :]$) with 3×3 average pooling kernels for neighborhood statistics.

Absorption component ($\mathcal{A}$) quantifies the L2 distance between low-frequency features, modeling how wavelength-dependent color attenuation leads to significant absorption effects in underwater imaging [170].

$$\mathcal{A}(\mathbf{y}, \mathbf{z}) = \|\mathbf{f}_{LF}(\mathbf{y}) - \mathbf{f}_{LF}(\mathbf{z})\|_2, \tag{4.21}$$

where $\mathbf{f}_{LF} = \mathbf{f}_{DE}[:, :, : C/2]$ represents components primarily affected by underwater absorption.

To adaptively balance scattering and absorption, D2Mamba introduces a gating function $\omega(\mathbf{y})$, which is based on a two-layer MLP with ReLU activation. This allows the heuristic to prioritize scattering or absorption cues depending on local feature characteristics. By integrating these terms into the A* framework, GIFH enables paths to adapt to the physical properties of the input image, moving beyond rigid raster or zig-zag scans.

4.2.1.5 Gated Synthesis Decoder (GSD)

To reconstruct the final enhanced image, a Gated Synthesis Decoder (GSD) is designed that adaptively fuses the corrected latent features from the Mamba block (f_M) with the original encoder features (f_{DE}). Instead of simple addition or concatenation, GSD learns a pixel-wise gating map, allowing it to selectively balance corrected features from the Mamba block with structural details from the encoder.

$$\begin{aligned} \mathbf{f}_{\text{fused}} &= (1 - \mathbf{G}) \odot \mathbf{f}_{DE} + \mathbf{G} \odot \mathbf{f}_M, \\ \mathbf{G} &= \sigma(\text{Conv}_{3\times3}([\mathbf{f}_{DE}; \mathbf{f}_M])), \end{aligned} \tag{4.22}$$

where $\odot$, $\mathbf{G}$, σ, and $[\cdot;\cdot]$ represent element-wise multiplication, a spatially-adaptive gate map, the sigmoid activation function and channel-wise concatenation, respectively.

The fused representation is refined using a series of transpose convolutional layers and ReLU activations to reconstruct the output image (I_E). This method enhances global adjustments from the Mamba block with local details from the encoder, balancing corrections and features for sharper, artifact-free, and perceptually consistent underwater image enhancement.

4.2.1.6 Training Losses

This work adopts mean squared error loss (L_2) [171] for pixel-level reconstruction, perceptual loss (L_P) [122] for feature-level similarity, and structural similarity loss (L_{SSIM}) [171] for preserving structural details of the proposed framework, which plays a crucial role in training the model for accurate underwater image enhancement. In addition to these, a Spectral

	Methods	DWNet TOMM'23	U-Shape TIP'23	DM-Water MM'23	Unformer TAI'24	Spectroformer WACV'24	CE-VAE WACV'25	Phaseformer WACV'25	SS-UIE AAAI'25	Proposed
	$\#Params(M)\downarrow$	0.48	65.6	10.00	7.31	2.40	80.74	1.77	4.25	**0.78**
	$Flops(G)\downarrow$	18.2	66.2	-	13.98	15.75	245.92	13.00	19.37	**7.06**
UIEB	SSIM ↑	0.843	0.910	0.909	0.861	0.917	0.870	0.928	0.909	**0.937**
	PSNR ↑	23.17	22.91	23.03	23.63	24.96	23.55	25.98	25.04	**25.70**
	UIQM ↑	4.237	4.330	2.908	4.238	4.200	4.323	4.418	4.441	**4.495**
	UISM ↑	7.236	7.296	4.058	7.271	7.312	7.299	7.132	7.086	**7.378**
LSUI	SSIM ↑	0.924	0.932	0.919	0.910	0.885	0.902	0.922	0.891	**0.935**
	PSNR ↑	28.98	26.13	29.56	28.88	25.08	29.65	29.60	28.61	**29.69**
	UIQM ↑	4.554	4.381	-	4.383	4.265	4.442	4.400	4.645	**4.669**
	UISM ↑	7.122	6.927	-	7.045	6.938	6.964	7.020	6.941	**7.144**
U45	UIQM ↑	4.549	3.997	3.854	4.133	3.921	4.007	4.245	4.250	**4.617**
	UISM ↑	6.873	6.962	7.120	7.005	7.088	6.959	7.099	7.073	**7.270**
	BRISQUE ↓	20.986	21.565	19.334	27.068	18.728	24.520	17.504	19.775	19.969
UCCS	UIQM ↑	4.476	4.201	4.301	4.436	4.229	4.352	3.980	4.278	**4.520**
	UISM ↑	6.988	6.913	7.008	6.843	6.898	6.887	6.974	6.869	**6.989**
	BRISQUE ↓	26.977	28.073	26.429	28.682	26.392	27.933	27.710	30.129	**25.832**
C60	UIQM ↑	4.227	4.514	4.184	4.204	4.158	4.375	4.286	4.417	**4.688**
	UISM ↑	7.438	6.764	7.421	7.299	7.488	7.353	7.499	7.422	**7.476**
	BRISQUE ↓	25.764	31.183	23.709	29.738	24.951	27.515	23.168	26.509	25.170
SQUID	UIQM ↑	2.434	2.266	-	-	2.558	1.965	2.335	2.135	**2.557**
	UISM ↑	7.373	7.275	-	-	7.193	7.189	7.357	7.302	**7.323**
	BRISQUE ↓	17.468	27.919	-	-	26.593	33.291	29.196	21.884	**11.489**

Table 4.10: *Quantitative comparison of D2Mamba with existing methods on two full-reference datasets (UIEB, LSUI) and four no-reference datasets (U45, UCCS, C60, SQUID). The top-performing method is highlighted in red, the second-best in blue, and the third-best in green. (↑ higher is better, ↓ lower is better).*

Wasserstein Attenuation Loss (L_{SWAL}) is designed to enforce distributional alignment in the spectral domain. The total loss function is formulated as follows:

$$L_T = \omega_1.L_2 + \omega_2.L_P + \omega_3.L_{SSIM} + \omega_4.L_{SWAL}, \tag{4.23}$$

where, $\omega_{1,2,3,4} \in \{1, 0.02, 0.5, 0.1\}$ are empirically determined weighting factors.

Spectral Wasserstein Attenuation Loss (SWAL). Underwater light attenuation follows an exponential decay model that is spectrally dependent:

$$I_c(x) = J_c(x)\, e^{-\beta_c d(x)} + B_c \left(1 - e^{-\beta_c d(x)}\right), \tag{4.24}$$

where $I_c(x)$ is the observed intensity at pixel x in channel $c \in \{R, G, B\}$, $J_c(x)$ is the scene radiance, $d(x)$ is the scene depth, β_c is the attenuation coefficient, and B_c is the background light. Most existing loss functions treat channels independently or apply heuristic per-channel penalties, which fails to capture the continuous spectral nature of attenuation.

To address this, the SWAL enforces distributional alignment between restored images and natural clean images in a perceptually uniform color space (CIE Lab). Formally, let P_{pred} denote the distribution of pixel colors in the enhanced image, represented in Lab space, and let P_{ref} denote the distribution of natural clean image colors. In this work, P_{ref} is approximated using large-scale natural image statistics from ImageNet [172]. SWAL is defined as the Wasserstein-1 distance [173] between the two distributions:

$$L_{\text{SWAL}} = \inf_{\gamma \in \Pi(P_{pred}, P_{\text{ref}})} \mathbb{E}_{(x,y)\sim\gamma}\big[\|x - y\|\big], \tag{4.25}$$

where $\Pi(P_{pred}, P_{\text{ref}})$ is the set of joint distributions with marginals P_{pred} and P_{ref}. For computational efficiency, SWAL is approximated using the Sliced Wasserstein Distance (SWD) with K random projections $\theta \in \mathbb{S}^{d-1}$:

$$L_{\text{SWAL}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\big|X_{\theta_k}^{(i)} - Y_{\theta_k}^{(i)}\big|, \tag{4.26}$$

where $X_{\theta_k}^{(i)}$ and $Y_{\theta_k}^{(i)}$ are the sorted projections of predicted and reference samples along direction θ_k. By aligning enhanced images with perceptually and physically consistent color distributions, SWAL mitigates unnatural color shifts (e.g., over-saturated reds) and constrains the enhancement process to remain consistent with underwater light attenuation, all without requiring adversarial supervision.

4.2.2 Experiments

4.2.2.1 Implementation Details

The proposed model was implemented in PyTorch 2.8 and trained and tested on an NVIDIA A100 GPU. The AdamW optimizer was adopted with an initial learning rate of 0.0005, and input images were resized to 256 × 256 pixels with a batch size of 5. The training was conducted for 150 epochs, using L2 loss, SSIM loss, perceptual loss, and SWAL loss to balance pixel-level accuracy and perceptual quality.

4.2.2.2 Datasets

This work utilized two real-world datasets for training and evaluation: UIEB [30] and LSUI [74]. To further assess the generalization capabilities of the proposed model in practical scenarios, evaluated it using the U45 [125], C60 [30], UCCS [124], and SQUID [174] benchmarks that do not have ground-truth references.

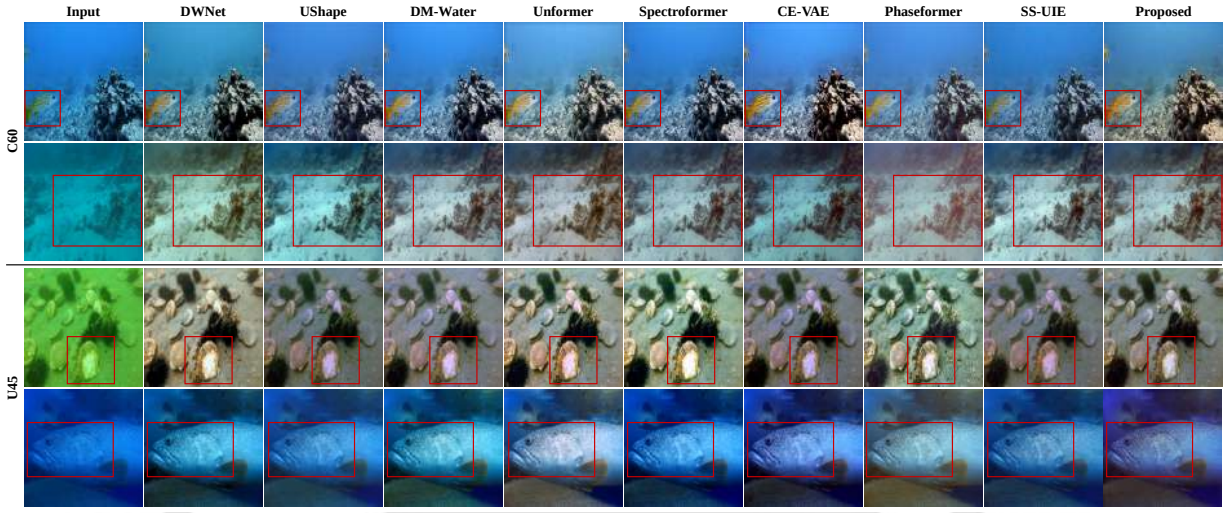


Figure 4.21: *Qualitative results on C60 and U45. D2Mamba method produces clearer structures and more natural colors compared to other existing methods.*

4.2.2.3 Comparison Methods

The proposed model is compared against several SOTA UIE methods: DWNet [72], UShape [74], DM-Water, Unformer [148], Spectroformer [77], CE-VAE [175], Phaseformer [176], and SS-UIE [104]. All methods are implemented using the authors' official source codes and evaluated under identical experimental settings for fair comparison.

4.2.2.4 Evaluation Metrics

Th proposed method is evaluated using both full-reference and no-reference metrics. PSNR and SSIM measure pixel-level fidelity and structural similarity against ground truth [114], while UIQM and UISM provide underwater-specific perceptual quality assessment [117]. Additionally, BRISQUE [118] is included as a general no-reference metric, where lower scores indicate more natural and less distorted images.

4.2.2.5 Quantitative Evaluation

Table 4.10 compares D2Mamba with recent SOTA UIE methods across two benchmark datasets. On the LSUI dataset, which features diverse degradation patterns, D2Mamba outperforms all models, achieving an increase of 0.04 dB in PSNR and 0.32% in SSIM, indicating its effectiveness in recovering fine details. For the UIEB dataset, focused on real-world underwater images, the proposed method shows substantial gains, with a 0.97% increase in SSIM and a 1.22% gain in UIQM, demonstrating more visually pleasing underwater images, especially with natural colors and improved contrast. These consistent improvements across datasets underscore D2Mamba's ability to handle different underwater environments while ensuring pixel-level accuracy and perceptual realism. Apart from that, in non-reference evaluations D2Mamba achieves the best overall performance across multiple datasets. On UCCS and U45, the proposed method ranks first or second in most metrics but secures the top score in UIQM, with overall gains of 1.49% and 0.98% over the best existing methods. On SQUID, D2Mamba remains highly competitive, with only a 0.001 gap from the strongest baseline on UIQM. These results confirm the robustness of D2Mamba under real-world conditions without ground-truth references. D2Mamba has only 0.78 million parameters and 7.06 GFLOPs, making it lightweight and useful for resource-constrained systems.

4.2.2.6 Qualitative Evaluation

Figure 4.21 presents visual comparisons across several challenging underwater scenes. Competing methods often suffer from color distortion, residual haze, or over-enhancement artifacts. For instance, U-Shape and DM-Water tend to produce images with noticeable bluish or greenish tones, while Spectroformer and CE-VAE frequently lead to over-saturation. SS-UIE and Phaseformer improve contrast but still leave haze in low-visibility regions. DWNet produces sharper edges but struggles to recover natural colors. In contrast, the proposed method restores both color fidelity and structural details, yielding clearer object boundaries, natural tones, and balanced contrast. These visual improvements confirm the quantitative gains, demonstrating the robustness and perceptual advantage of D2Mamba across diverse

L_2	L_{SSIM}	L_P	L_{SWAL}	SSIM	PSNR
✓	×	×	×	0.918	24.88
✓	✓	×	×	0.931	25.14
✓	✓	✓	×	0.934	25.26
✓	✓	✓	✓	**0.937**	**25.70**

Table 4.11: *Comparison of quantitative results on D2Mamba obtained using different loss settings on the UIEB dataset.*

Scan Type	Latent Pixels	SSIM	PSNR
Raster	4096	0.926	25.30
Bidirectional	4096	0.928	25.37
Cross-scan	4096	0.929	25.43
Diagonal	4096	0.928	25.38
GIFH guided A^*	**110 - 125**	**0.937**	**25.70**

Table 4.12: *Comparison of quantitative results on D2Mamba using various scanning types on the UIEB dataset.*

underwater conditions.

4.2.2.7 Ablation Study

This ablation study explores the effectiveness of each component in the proposed framework, providing valuable insights into their roles in the overall design.

Effect of scan strategies. The A^* pathfinding strategy significantly reduces computational load when applied to the UIEB dataset, which was used solely for ablation testing. At a feature dimension of 64×64, it processes only 110 to 125 latent pixels per underwater image, compared to 4,096 for traditional methods, resulting in a $\sim$ 97% reduction in complexity. This efficiency leverages spatial correlations in underwater degradation and real-

$\#Paths$	$\#MamabaLayers$	SSIM	PSNR
8	2	0.937	25.68
4	**4**	**0.937**	**25.70**
2	8	0.932	25.27

Table 4.13: *Comparison of quantitative results on D2Mamba by varying path numbers and Mamba layers on the UIEB dataset.*

time capabilities essential for underwater robotics while maintaining enhancement quality by processing only about 3% of the pixels directly. The qualitative measures of the ablations can be found in Table 4.12.

Effect of number of paths in A^* and Mamba Layers. As shown in Table 4.13, balancing paths and layers yields the best performance. The (4,4) setting achieves the highest PSNR (25.70) and maintains the best SSIM (0.937), while skewed designs with many paths or many layers lead to equivalent or less results. This highlights the significance of structural balance in Mamba-based framework (refer to Figure 4.23).

Effect of loss terms. The effect of different loss functions is evaluated on underwater image enhancement. The combination of pixel, perceptual, structural, and the proposed spectral loss leads to clear improvements in both quantitative metrics and visual quality. As shown in Table 4.11 and Figure 4.22, incorporating SWAL further enhances color consistency and realism, validating its effectiveness.

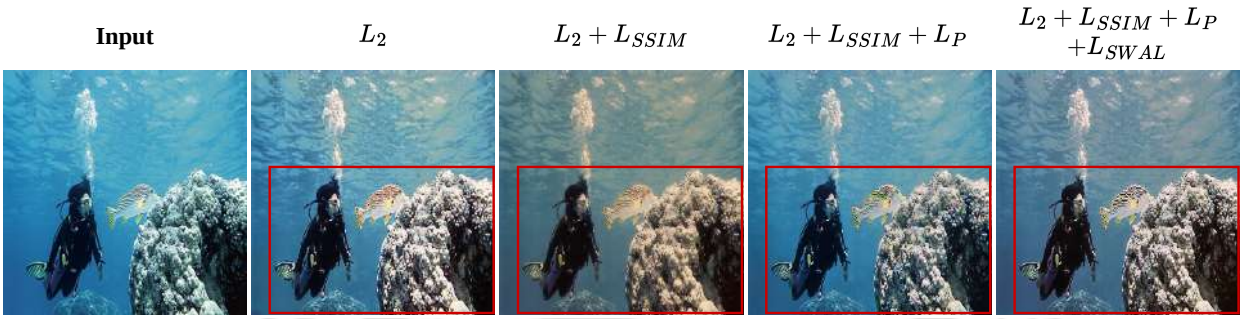


Figure 4.22: *Enhanced image under various loss configuration.*

4.3 Summary

This contributory chapter presented details of two models (UEnhancer and D2Mamba). In UEnhancer, this work proposes a novel transformer-based architecture for enhancing underwater images. The primary innovation lies in the incorporation of dual color spaces, allowing for the effective learning of a diverse range of discriminative features. Additionally, the SAE block is introduced, along with modifications to the Multi-Head Attention (MHA)

Figure 4.23: *Visualization of learned paths using GIFH.*

and Feed-Forward Network (FFN) blocks, to improve performance by better capturing local context and addressing limitations identified in previous research. Furthermore, the multi-scale frequency cube-based SSIM loss function is presented to ensure that UEnhancer produces images with greater precision and detail. Experimental results demonstrate that UEnhancer surpasses existing techniques across various challenging scenarios. The visual quality of enhanced images shows a significant improvement, exhibiting more realistic colors and enhanced visibility. An ablation study is also conducted to clarify the contributions of different architectural modules and parameter settings. Moreover, UEnhancer notably enhances the quality of low-light images and proves effective for high-level underwater applications.

In the second work, D2Mamba, a lightweight Mamba-based model, is proposed for enhancing ocean-water images. The model integrates dual-stream spatial-frequency encoding, geodesic-guided feature traversal, Mamba-based correction, and gated synthesis reconstruction to capture local details and global dependencies effectively. Further, a Spectral Wasserstein Attenuation Loss enforces spectral consistency, enabling perceptually accurate color restoration. D2Mamba outperforms existing methods with lower computational requirements, achieving a strong balance of efficiency, robustness, and visual fidelity.

While the previous approaches improved performance for typical types of degradation present in the dataset, real underwater environments often exhibit multiple degradations simultaneously, such as haze, low contrast, blur, and color distortion. Models designed for a single type of degradation often struggle with diverse conditions, which limits their effectiveness in real scenarios. To address this issue, the next contributory chapter will present a unified multi-degradation enhancement model that can handle various underwater degradations simultaneously.

The second work (D2Mamba) presented in this chapter has been accepted for publication at the WACV 2026 conference.

"You can't cross the sea merely by standing and staring at the water."

~Rabindranath Tagore

5

Multi-degradation Learning Model for Ocean-water Image Enhancement

While significant progress has been made in the enhancement of underwater images [74, 177, 178], multi-degraded UIE is still an open problem. The current methods either focus on correcting color distortion [35], removing haze [36], and enhancing image contrast [37], or they use separate models trained to address specific degradations. Typically, these models are organized in a sequence, where one process follows another, for instance [62] addressed color correction followed by haze removal. However, this process may lead to unpleasant outcomes, due to (1) dependency of outputs from the first stage on the following stages, and (2) necessity of a degradation-specific model. Thus, the enhancement model must perform

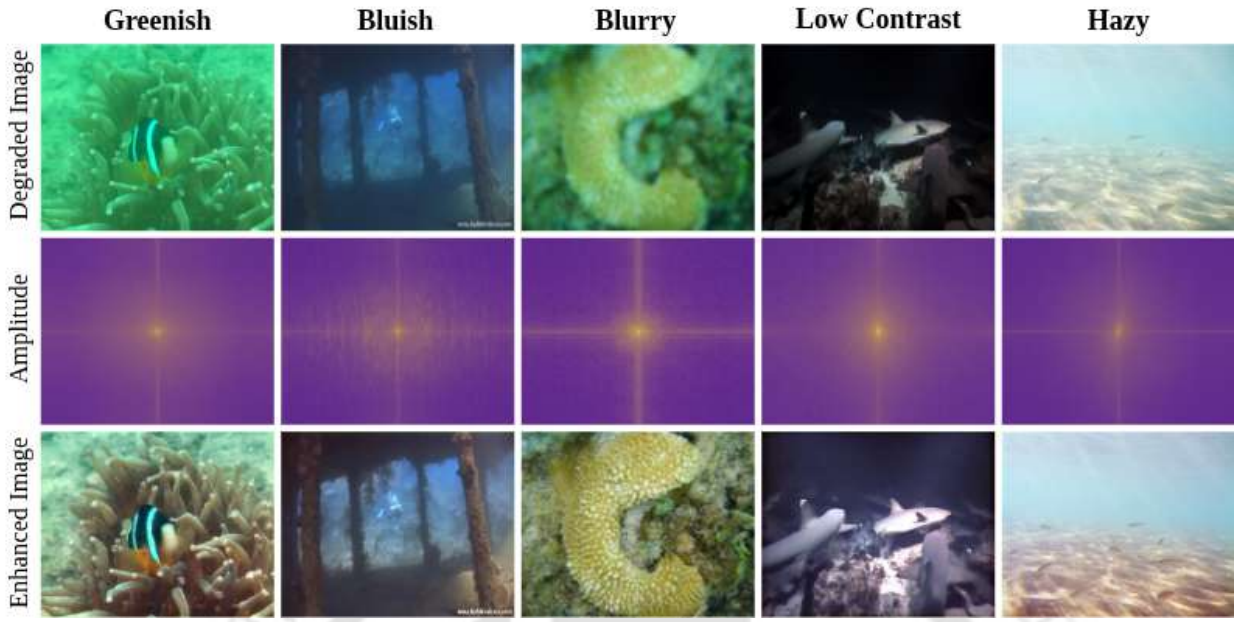


Figure 5.1: *Degraded images, Fourier spectra of degraded images, and enhanced images by PLUM.*

well under all degradation conditions.

To tackle the challenges mentioned, this work proposes a unified approach to enhance images affected by multiple types of degradation. In order to accomplish this, it initially employs a domain-translation technique that allows the model to convert any degraded image into a particular domain (color distortion, hazy, blurry, and low contrast). These variations help the model to understand a collection of characteristics that describe all degraded images. Then, an enhancement network is proposed that inputs domain-translated and original degraded images to generate degradation-free output. Fig. 5.1 shows the visual

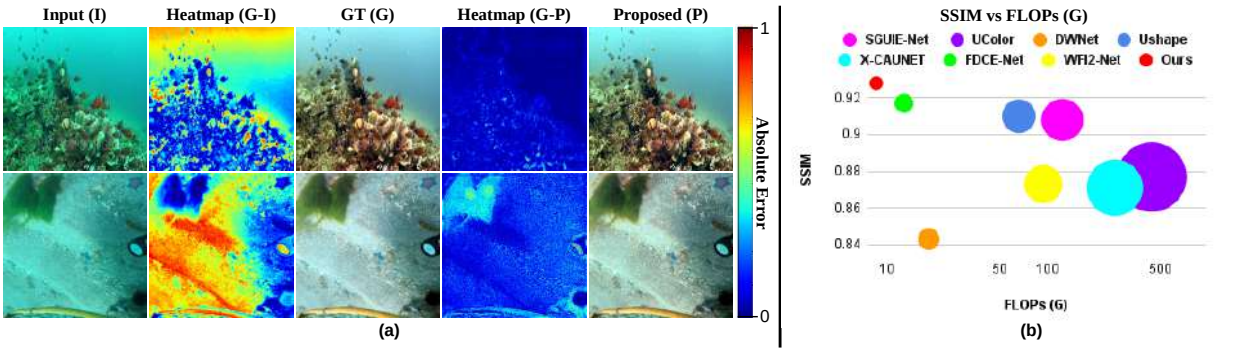


Figure 5.2: *(a) The heatmap shows the pixel-wise absolute error between the input image and the ground truth, as well as the enhanced image and the ground truth. (b) Trade-off between SSIM vs FLOPs (G) on UIEB dataset.*

example of input, amplitude, and enhanced image. It can be seen that different degradations pay different attention to the amplitude spectrum. Therefore, integrating the frequency domain with the spatial domain improves or suppresses specific features, enabling effective enhancement in challenging underwater conditions. Recently, restoration techniques utilized prompt guidance [179–181] on generating visual prompts from training datasets to enhance the model's learning capabilities. Therefore, it aims to leverage the capabilities of learnable prompts in both frequency and spatial domains to enhance the multi-degraded images. Fig. 5.2(a) and (b) show the pixel-wise absolute error through heatmap, along with the trade-off between SSIM and FLOPs (G). In this map, cooler colors (blue) represent lower reconstruction errors and better enhanced quality, while warmer colors (yellow to red) indicate higher errors, showing where the model's output deviates from the ground truth. The proposed model outperforms the state-of-the-art methods. The major contributions are:

- A novel lightweight unified model for enhancing multi-degraded underwater images. It employs a single trainable model to address all types of degradation.
- Developed Spatial-Fourier Mutual Learning (SFML) block to effectively investigate features across spatial and frequency domains by incorporating Fast Fourier Transform, enabling a more comprehensive understanding of image representations.
- Designed a Hierarchical Cross-encoder Feature Swapping (HCFS) strategy that enhances multi-degraded domain-invariant representation learning, improving generalization and enhancement quality.
- Studied how prompts can improve the interaction between the spatial and frequency domains by using cross-attention, allowing for better feature interaction and representation.

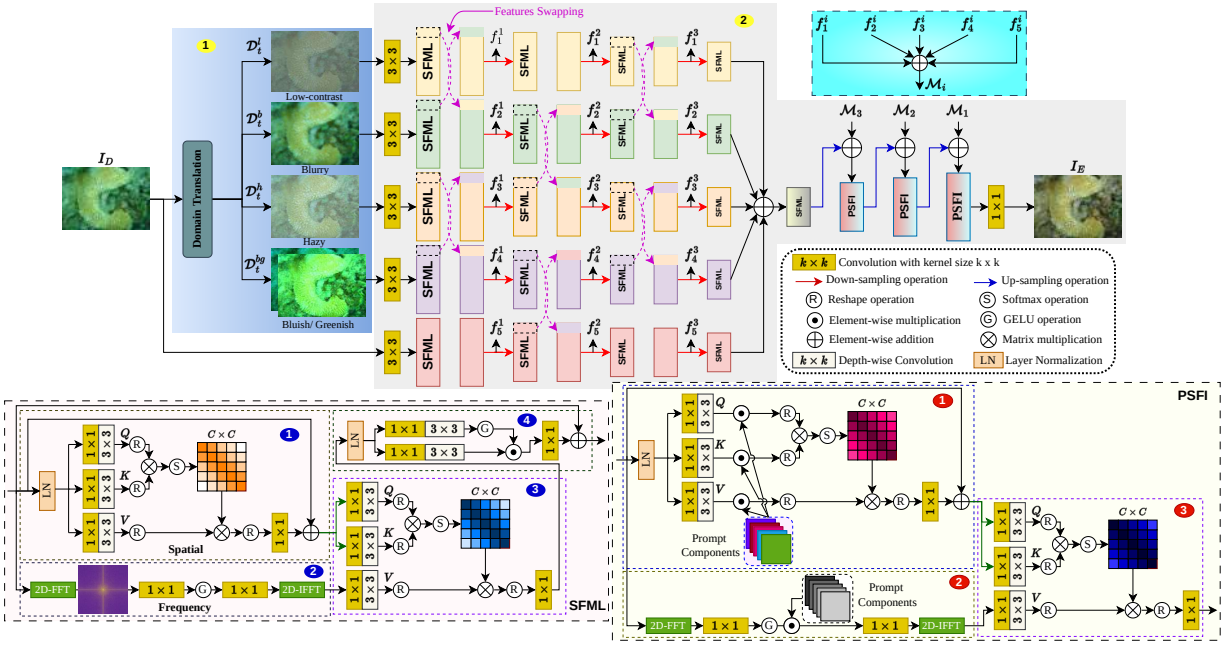


Figure 5.3: *Overview of the proposed model with all the sub-components.*

5.1 Proposed Method

This work proposes a solution for addressing multiple degradations using a unified model. It incorporates several key designs to enhance a degraded underwater image in the proposed model. First, it briefly describes the overall pipeline of the proposed model as depicted in Fig. 7.2. Then, it introduces three components of the proposed model: Domain translation, Spatial-Fourier mutual learning (SFML), Hierarchical cross-encoder feature swapping (HCFS), and prompting-based Spatial-Fourier interaction (PSFI).

Overall Pipeline. An overview of the proposed model is illustrated in Fig. 7.2. Given a degraded underwater image ($I_D \in \mathbf{R}^{3\times h\times w}$), the goal is to generate its different degraded versions using domain translation. For instance, this work transforms a degraded underwater image $I^D \in \mathbf{R}^{3\times h\times w}$ into a low-contrast ($D_t^l \in \mathbf{R}^{3\times h\times w}$), blurry ($D_t^b \in \mathbf{R}^{3\times h\times w}$), hazy ($D_t^h \in \mathbf{R}^{3\times h\times w}$), and color-distorted ($D_t^{bg} \in \mathbf{R}^{3\times h\times w}$) version. These inputs are passed through the five different convolution layers of kernel size of 3×3 to get the shallow features. Then, these features are fed in parallel into five different three-level encoder blocks to learn multiple degraded features. At each level, feature sharing is performed between different encoder layers to learn the domain-relevant features effectively. Each encoder block helps the

network better characterize the internal relationship in the spatial and frequency domains by introducing the Fast Fourier transform in the network, namely, spatial-Fourier mutual learning (SFML) block. After the final encoder layer, the learned features are aggregated and processed through successive decoder layers for reconstruction. Since the clean image remains common across degradations, the extracted feature suppresses degradation-specific details, focusing on essential features. During reconstruction, the learned encoder features are summed up and sent to the corresponding decoder through a skip connection. Each decoder layer contains the proposed prompt-guided spatial-frequency interaction (PSFI) block to transform the abstract representation of the encoder into meaningful features. Finally, this work applies a 1×1 convolution to the decoded features to obtain the enhanced image $I^E \in \mathbf{R}^{3 \times h \times w}$.

5.1.1 Domain Translation (DT)

The domain translation method takes a degraded image as input and converts it to other domains. This work converted a degraded input into low-contrast, blurry, hazy, and bluish-greenish domain.

Low-contrast [182] image generation involves reducing the intensity differences between pixels, making the image appear faded or dull. This is achieved by scaling pixel values toward the mean intensity, effectively compressing the dynamic range. The low-contrast transformation is applied as

$$D_t^l = \alpha I^D + (1 - \alpha) M. \tag{5.1}$$

where M is the mean intensity, and α controls the contrast level. The α value is set randomly, this makes low-contrast images more diverse and realistic, improving generalization, robustness, and real-world adaptability.

To transfer into the **blurry** domain [182], a Gaussian function is used that weights neighboring pixels based on their distance from the center (refer to eq. 5.2 and 5.3). Closer pixels receive higher weights, leading to softened edges and reduced image sharpness. The blur intensity can be adjusted by changing the kernel size and the standard deviation (sigma);

larger values result in a stronger blur. In this instance, the kernel size is set to 5 and the sigma value to 1.0.

$$D_t^b(i,j) = \sum_{x=-k}^{k} \sum_{y=-k}^{k} I^D(i+x, j+y)G(x,y). \tag{5.2}$$

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-(\frac{x^2+y^2}{2\sigma^2})}. \tag{5.3}$$

Hazy image generation [182] simulates real-world atmospheric scattering by blending an image with an artificial haze effect. This process utilizes a haze formation model (refer to eq. 5.4), where the resulting hazy image is a combination of the original image and atmospheric light (A_L), modified by a transmission map (t_m). By adjusting parameters such as t_m and A_L, different levels of haze can be applied.

$$D_t^h = I_D t_m + A_L(1 - t_m). \tag{5.4}$$

In this model, rather than using an exponential decay, it randomly assigns the transmission value t_m between 0.5 and 1.0. This approach results in haze thickness that is uniformly random. The atmospheric light A_L is set to $(0.8, 0.8, 0.8)$ for the RGB channels, which reflects a bright grayish haze. By introducing this randomness, it enhances the model's generalizability to real-world hazy images.

Translating a **bluish or greenish** image domain involves manipulating the color balance of an image to emphasize specific hues. For a bluish image, the blue channel is heightened while reducing the intensity of the red and green channels, resulting in a cool-toned appearance with dominant shades of blue. Similarly, the green channel is enhanced for a greenish image, and the red and blue channels are reduced, giving the image a fresh, nature-inspired feel with lush green tones.

$$D_t^{bg} = Concat(ColorTrans^b(I^D), ColorTrans^g(I^D)). \tag{5.5}$$

5.1.2 Spatial-Fourier Mutual Learning

Enhancing a degraded image with multiple degradations requires analyzing features beyond the spatial domain, as this may not capture global details and could lead to unwanted artifacts. In contrast, the SFML block is designed by utilizing the Fast Fourier Transformation to analyze the features in the frequency domain along with the spatial domain. Learning from both simultaneously improves haziness, blurriness, color distortion, and uneven illumination, producing more precise and visually appealing underwater images. Additionally, mutual learning between spatial and frequency branches enables efficient feature extraction, reducing redundant processing and enhancing model generalization across different degradations. As shown in Fig. 7.2, SFML has four parts. First, it takes input and performs multi-head self-attention in the spatial domain to obtain F_s.

$$F_s = MHSA(LN(x)) + x. \tag{5.6}$$

Second, it applies the 2D-FFT on the input and analyzes the features in the frequency domain to get F_f.

$$F_f = IFFT(c^{1\times1}(GELU(c^{1\times1}(FFT(x))))). \tag{5.7}$$

Third, through the cross-attention mechanism, both spatial and frequency domain features are learned mutually, and it enhances the contextual understanding and enables better feature fusion, making it valuable in UIE tasks.

$$F_{ca} = Cross_Attn(F_s, F_f). \tag{5.8}$$

Fourth, these fused features are further refined with a gated feed-forward network (GF). Let SFML take x as input, then

$$F_{SFML} = GF(F_{ca}) + x. \tag{5.9}$$

5.1.3 Hierarchical cross-encoder feature swapping (HCFS):

As shown in Fig. 7.2, the proposed model comprising five parallel encoders, each with three hierarchical levels, features swapping across corresponding levels introduces a powerful mechanism that improves the model's representation abilities through structured feature interaction. At each semantic level, features from a few encoders are selectively exchanged, enabling the model to aggregate complementary information extracted under varied domain-specific priors. This design mitigates redundancy among encoders and leverages their diversity to construct a more expressive joint latent space. As a result, the model gains improved robustness to noise and varying image artifacts. It is particularly beneficial when dealing with multi-degradation contexts, making it highly effective for complex underwater image enhancement. Let $F_{e1} \in \mathbf{R}^{c\times h\times w}$ and $F_{e2} \in \mathbf{R}^{c\times h\times w}$ be the features of two different encoders, then the swapped features:

$$F_{e1}^{s} = \begin{cases} F_{e2}[c,:,:] & \text{if } c < c_s, \\ F_{e1}[c,:,:] & \text{otherwise} \end{cases} \tag{5.10}$$

$$F_{e2}^{s} = \begin{cases} F_{e1}[c,:,:] & \text{if } c < c_s, \\ F_{e2}[c,:,:] & \text{otherwise} \end{cases} \tag{5.11}$$

where $c_s = [\alpha.c]$ indicates the number of channels to swap, α indicates the fraction of channels to swap. This work deterministically swaps the top 12.5% of channels ($\alpha = 1/8$) between paired encoder features at each level. This selective fusion enables partial feature alignment while preserving most of the encoder's native representation.

5.1.4 Prompting-based Spatial-Fourier Interaction

Prompt-based Spatial-Frequency Interaction (PSFI) enhances multi-degraded underwater images by leveraging spatial and frequency domain features through learned prompts. This method allows the model to focus on important details, improving contrast, texture clarity,

and structural consistency while minimizing artifacts. In the PSFI module, visual prompts ($V_p \in \mathbf{R}^{c\times h\times w}$) are initialized as learnable parameters, similar to prompt tokens in prompt-based vision models [180, 181]. Specifically, it uses a set of trainable embeddings randomly initialized and jointly optimized with the rest of the network.

$$V_p = [p^1, p^2, p^3, \ldots, p^c], \quad p^i \in \mathbb{R}^{h\times w}. \tag{5.12}$$

PSFI first applies spatial prompting for better interpretability, directing focus to meaningful regions to get the attentive features F_{ps}, and then incorporates prompt components in the frequency domain to address specific degradations like blur and noise to obtain refined features F_{pf}.

$$\begin{aligned} Q^p &= Q \odot V_p, \quad K^p = K \odot V_p, \quad V^p = V \odot V_p, \\ F_{ps} &= (\text{MHSA}(Q^p, K^p) \otimes V^p) + x, \\ F_{pf} &= \text{IFFT}(c^{1\times1}(\text{GELU}(c^{1\times1}(\text{FFT}(x))) \odot V_p)). \end{aligned} \tag{5.13}$$

Finally, by jointly learning from both spatial and frequency features through cross-attention, the model becomes more adaptable to real-world UIE, where spatial and frequency distortions vary significantly. The overall process of PSFI is defined as:

$$F_{PSFI} = Cross\_Attn(F_{pf}, F_{ps}). \tag{5.14}$$

5.1.5 Loss Function

Using a similar strategy as [73], the color-channel specific L1 loss ($\mathcal{L}_1^c$) is utilized to supervise the generated image in the proposed model. Additionally, the perceptual loss ($\mathcal{L}_p$) is computed using the pre-trained VGG-16 network ($\phi(.)$), which has been trained on the ImageNet dataset. The overall loss function of the proposed model can be expressed as

follows:

$$\begin{aligned}
\mathcal{L}_T &= \mathcal{L}_1^c + \lambda \mathcal{L}_p, \quad \text{where} \\
\mathcal{L}_1^c &= \sum_{c \in \{r,g,b\}} w_c \left(\frac{1}{M} \sum_{i=1}^{M} \left\| (I_i^E)_c - (I_i^g)_c \right\|_1 \right), \\
\mathcal{L}_p &= \frac{1}{M} \sum_{i=1}^{M} \left\| \phi(I_i^E) - \phi(I_i^g) \right\|_2^2.
\end{aligned} \tag{5.15}$$

where w_c is the weight factor for different color channels and it is set to 1, 1.5, and 2, for R, G, and B channels, respectively. λ is deciding the importance of perceptual loss and it is set to 0.5. I^E and I^g are generated and ground truth image.

5.2 Experiments

5.2.1 Dataset Details

This work is trained and tested using the real-world UIEB [30], LSUI [74], and SUIM-E [75] datasets. Additionally, to assess the generalization capabilities, non-reference benchmarks C60 [30], U45 [125], and UCCS [124] are used for testing. These datasets provide a diverse range of marine organisms and environments for analysis.

5.2.2 Implementation Details

This model is developed using PyTorch 2.0.0 and conducted training and testing on an NVIDIA A100 GPU. The Adam optimizer was utilized, with parameters $\beta 1 = 0.9$ and $\beta 2 = 0.999$. The patch size is set to 256×256, the batch size to 5, and epoch to 300. The learning rate was kept at 0.0002. The convergence of the model's loss is presented in Fig. 5.8.

Table 5.1: *SSIM, PSNR, and UIQM result of the proposed method and SOTA methods on the UIEB dataset. Red, blue, and green indicates the 1^{st}, 2^{nd}, and 3^{rd} best results.*

Methods	Forums	SSIM	PSNR	UIQM
SGUIE-Net [75]	TIP'22	0.908	24.07	3.004
DWNet [72]	TOMM'23	0.843	23.17	2.897
Ushape [74]	TIP'23	0.910	22.91	2.725
Semi-UIR [177]	CVPR'23	0.901	24.59	-
X-CAUNET [73]	ICASSP'24	0.871	24.12	3.132
U-ENHANCE [76]	ACCVW'24	0.911	22.07	2.701
Spectroformer [77]	WACV'24	0.917	24.96	3.075
FDCE-Net [178]	CSVT'24	0.917	23.87	-
PhISH-Net [78]	WACV'24	0.869	23.43	-
WFI2-net [79]	CVPR'24	0.873	23.86	-
TPENet [83]	TIM'24	0.909	22.71	3.011
HG2former [93]	IJOE'25	0.890	23.82	-
UIEFORMER [92]	IJOE'25	0.906	23.17	-
UIE-SFIFormer [91]	IJOE'25	0.900	22.99	2.995
Proposed		**0.928**	**25.59**	**3.045**

Table 5.2: *SSIM and PSNR result of the proposed method and SOTA methods on the LSUI dataset.*

Methods	Forums	SSIM	PSNR
Water-Net [30]	TIP'19	0.82	18.73
FUnIE-GAN [32]	RA-L'20	0.84	22.63
Ucolor [33]	TIP'21	0.89	22.91
Semi-UIR [177]	CVPR'23	0.83	20.47
Ushape [74]	TIP'23	0.93	24.16
DM-water [132]	MM'23	0.89	27.65
FDCE-Net [178]	CSVT'24	0.88	24.47
WFI2-net [79]	CVPR'24	0.94	27.26
HG2former [93]	IJOE'25	0.90	26.77
UIE-SFIFormer [91]	IJOE'25	0.91	27.18
Proposed		**0.93**	**29.57**

5.2.3 Quantitative Results

To quantitatively evaluate the performance of the proposed UIE method, this work compared it using standard image quality metrics: PSNR, SSIM, and UIQM. Tables 5.1, 5.2, and 5.3 demonstrate that the proposed model consistently outperforms competing approaches. Specifically, compared to WFI2-Net, the proposed model shows a significant improvement of 1.73 dB on the UIEB dataset and 2.31 dB on the LSUI dataset, indicating superior detail preservation and structural fidelity. On the SUIM-E, it achieved an average PSNR of 26.65 dB and SSIM of 0.930, demonstrating that proposed approach more effectively restores fine details and reduces distortion compared to existing techniques. Additionally, Table 5.4

Table 5.3: *SSIM and PSNR result of the proposed method and SOTA methods on the SUIM-E dataset.*

Methods	Forums	SSIM	PSNR
FUnIE-GAN [32]	RA-L'20	0.825	23.590
SGUIE-Net [75]	TIP'22	0.857	25.987
DWNet [72]	TOMM'23	0.861	24.850
Ushape [74]	TIP'23	0.783	22.647
X-CAUNET [73]	ICASSP'24	0.886	24.721
Unformer [148]	TAI'24	0.879	25.097
Spectroformer [77]	WACV'24	0.889	25.473
Proposed		**0.930**	**26.65**

Table 5.4: *UIQM results of the proposed method and SOTA methods on the U45, C60, and UCCS datasets.*

Methods	WaterNet [30]	FUnIE-GAN [32]	Ucolor [33]	SGUIE-Net [75]	DWNet [72]	Ushape [74]	TPENet [83]	CBLA [183]	UIE-SFIFormer [91]	Proposed
U45	-	2.56	3.15	3.11	3.13	2.92	2.49	2.08	3.19	**3.21**
C60	2.38	2.51	2.48	2.53	2.27	2.76	-	2.12	2.74	**2.69**
UCCS	2.28	3.07	2.98	-	3.09	2.87	-	2.47	-	**3.03**

presents the UIQM comparison on the reference-less U45, C60, and UCCS datasets, where proposed model achieved results comparable to other models, demonstrating significant improvements in perceptual quality, especially in challenging conditions with poor visibility and strong color distortion.

5.2.4 Qualitative Results

Apart from the quantitative results, a qualitative evaluation was performed to assess the perceptual quality of the enhanced images. This evaluation aimed to capture visual fidelity, artifact presence, and naturalness, often not fully represented quantitatively. This work took four images from the UIEB and SUIM-E test set, including various degradation types. These images were processed using the proposed method and compared against existing methods. The qualitative results for UIEB are presented in Fig. 5.4. The proposed method shows competitive performance when compared to SOTA techniques. The proposed method effectively enhanced visibility in challenging conditions without introducing common artifacts such as over-saturation or unnatural color shifts. Additionally, this work provides qualitative insights into the U45, UCCS, and C60 datasets, as illustrated in Fig. 5.5. These

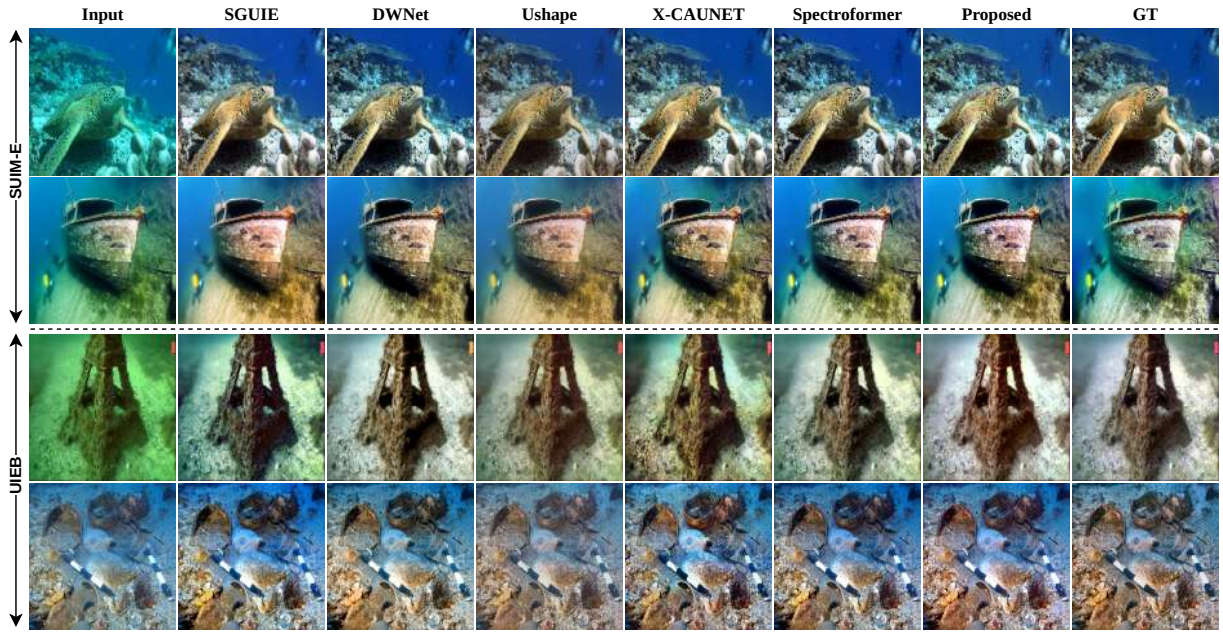


Figure 5.4: *The first and second row show the enhancement results of existing and the proposed methods on SUIM-E dataset, while the third and fourth row illustrate the enhancement results on UIEB dataset.*

findings highlight the substantial improvements in color correction and visibility of the enhanced images, which can be attributed to the innovative features presented in the proposed approach.

5.2.5 Ablation Study

In this section, an ablation study is conducted on UIEB dataset to validate the effectiveness of the different components.

- **Effect of loss function:** In this experiments, it evaluated various loss functions as outlined in Section 7.5. The results are summarized in Table 5.5. It is evident that the enhanced images produced by proposed model, which was trained using the complete loss function (i.e., the combination of the $\mathcal{L}_1^c$ loss and the $\mathcal{L}_p$ loss), show better quality compared to those generated by proposed model that did not incorporate the perceptual loss. This indicates the importance of including perceptual loss in order to enhance the visual quality of the final results, as illustrated in Fig. 7.3.

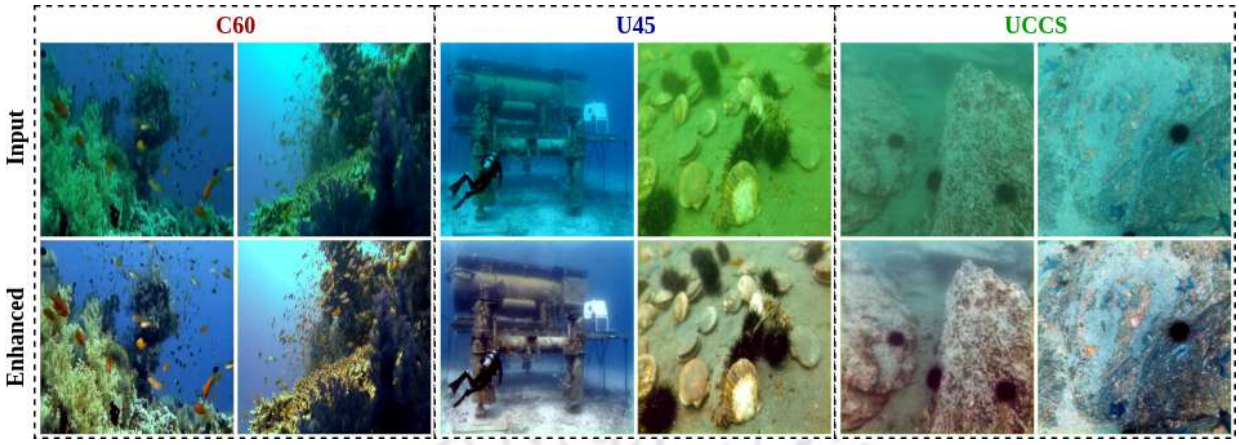


Figure 5.5: *Enhanced results on U45, C60, and UCCS dataset*

- **Effect of domain-translation:** To evaluate the effect of DT, it compared proposed model with and without DT using a cross-dataset approach. Table 5.6 shows the results of this setting. It is clearly seen that the incorporation of DT significantly improves the model's performance on cross-dataset.

- **Effect of feature swapping:** Swapping intermediate features between different degraded inputs enhances the model's performance, particularly in UIE as shown in the 2^{nd} row of Table 5.5. This aids in generalization and enables the model to better manage unseen real-world degradations.

- **Effect of prompting:** To examine the effectiveness of prompting, it conducted an ablation study, as shown in Table 5.5. The complete model demonstrated the highest performance, while disabling the prompt led to a noticeable decline in performance by 0.96 dB. These experimental results indicate that prompts significantly contribute to improving the quality of images.

- **Effect of Fourier domain analysis:** As demonstrated in Sec. 5.1.2 and 5.1.4, the absence of frequency domain information can hinder the model's ability to capture global details and may result in unwanted artifacts. Observing the results in the 1^{st} row of Table 5.6 and Fig. 7.3 reveals that incorporating the frequency information results in enhanced performance.

Table 5.5: *Quantitative result of different components of the proposed model on the UIEB dataset.*

Methods	SSIM	PSNR
w/o FFT	0.906	24.11
w/o feature swapping	0.905	24.35
w/o prompt	0.901	24.63
$\mathcal{L}_1^c$	0.911	24.22
Full Model ($\mathcal{L}_1^c + \lambda\mathcal{L}_p$)	**0.928**	**25.59**

Table 5.6: *Quantitative result of cross-dataset for UIE task.*

	Test →	LSUI		UIEB	
	Train ↓	SSIM	PSNR	SSIM	PSNR
w DT	**LSUI**	**0.93**	**29.57**	0.91	24.50
	UIEB	0.90	24.99	**0.93**	**25.59**
w/o DT	**LSUI**	**0.91**	**27.84**	0.89	23.09
	UIEB	0.88	22.97	**0.90**	**23.69**

5.2.6 Analysis of Parameters and FLOPs

This work assessed the parameters and FLOP counts of the UIE methods using an image size of 256 × 256. Table 5.7 shows that the proposed method is competitive regarding parameters and FLOP counts. The proposed model has 0.88 million parameters, making it competitive with existing deep learning-based UIE models and significantly lesser than Ucolor and Ushape. Despite using multiple encoders, feature swapping, and prompting, the model remains lightweight due to its efficient design and weight-sharing strategies. The proposed model requires approximately 8.59 GFLOPs per forward pass, balancing enhancement quality with real-time applicability.

5.2.7 Scalability Verification

A scalability assessment is performed to see how well the proposed model works with different amounts of computing power by changing the network's size and complexity. This work tested several setups, ranging from small to full-scale large versions. The larger (PLUM-L) version of the proposed model utilizes 16 feature channels in each convolution operation and features eight heads in the transformer block, while the PLUM-XL, PLUM-M, and PLUM-S versions contain 32, 12, and 8 channels, and 8, 6, and 4 heads, respectively. The evalua-

Figure 5.6: *Visual results of the ablation study.*

Table 5.7: *Parameters and GFLOPs comparison of the proposed method and SOTA methods.*

Methods	Parameters (M)	FLOPs (G)
FUnIE-GAN [32]	7.71	10.7
SGUIE-Net [75]	18.55	123.5
Ucolor [33]	157.4	443.9
DWNet [72]	0.48	18.2
Ushape [74]	65.6	66.2
DM-water [132]	10.0	-
X-CAUNET [73]	0.33	261.5
U-ENHANCE [76]	5.52	-
FDCE-Net [178]	5.53	12.8
WFI2-Net [79]	-	94.1
TPENet [83]	9	-
HG2former [93]	1.3	13.2
UIEFORMER [92]	1.7	96.96
Proposed	**0.88**	**8.6**

tion assessed two main aspects: performance, using image quality metrics (e.g., PSNR and SSIM), and efficiency, measured by FLOPs, parameter count, memory usage, and inference time. The results shown in Table 5.8 and Fig. 5.7 indicate that the proposed method remains competitive in performance even when reduced in size, making it ideal for environments with limited resources. Additionally, larger model variants exhibit improved enhancement quality, confirming that the model can effectively scale with increased capacity.

5.2.8 Convergence analysis

In the UIE task, training is performed using color-channel specific L1 loss and VGG (Perceptual) loss functions, each focusing on different aspects of image quality. The Fig. 5.8 illustrates how these losses behave during training. The observed loss trends indicate effective convergence:

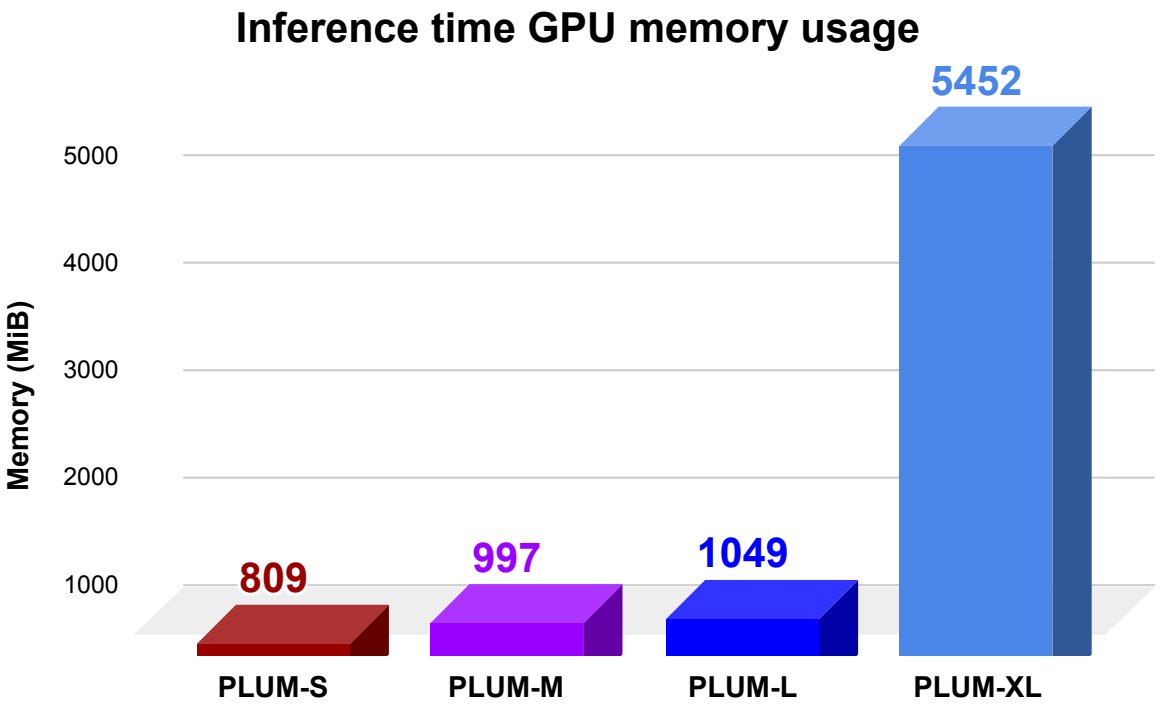


Figure 5.7: *Memory Usage of the proposed model from small to extra-large.*

Table 5.8: *Scalability Verification on the UIEB dataset.*

Methods	Params (M)	FLOPs(G)	Time (s)	SSIM	PSNR
PLUM-S	0.22	2.45	0.050	0.916	25.23
PLUM-M	0.50	5.05	0.060	0.922	25.17
PLUM-L	0.88	8.59	0.070	0.928	25.59
PLUM-XL	3.49	32.23	0.090	0.929	25.73

- The initial sharp drop indicates quick learning occurring in pixel and feature domains.
- A noticeable drop in all losses around early epoch due to better feature alignment after sufficient early training.
- The consistent, stable tail of the curves indicates that the model has reached a plateau and maintains consistent performance.

Overall, integrating color-channel-specific L1 and perceptual loss allows the model to reconstruct images that maintain structural integrity and visual coherence.

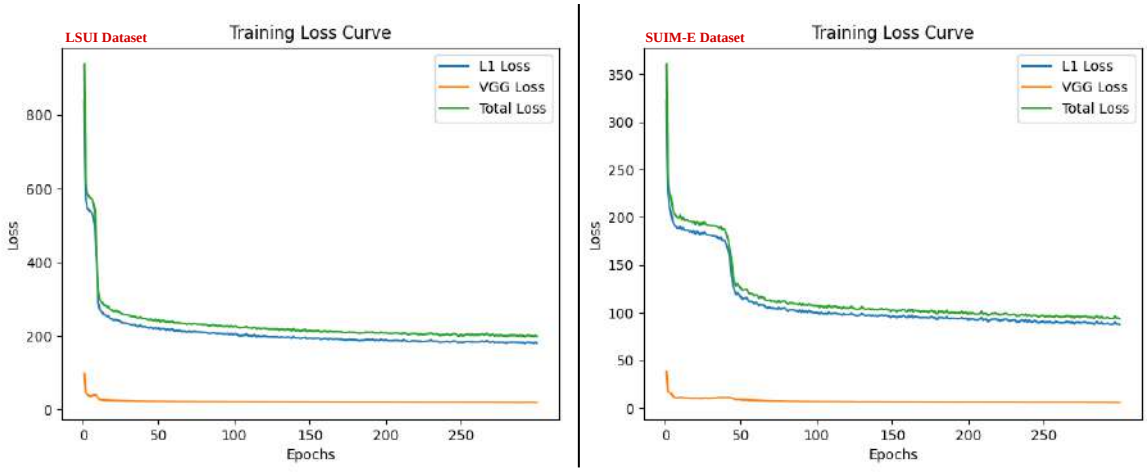


Figure 5.8: *Convergence of the loss curve.*

5.2.9 Application Test

This work validates the necessity and significance of the proposed method as a preprocessing step for underwater edge detection and depth estimation. As illustrated in Fig. 5.9, the proposed method achieves better visual outcomes and significantly improves edge detection; specifically, the structure of fish and human diver are preserved in edge detection, even outperforming ground truth. In the case of depth estimation, it can be seen that the enhanced image by the proposed model maintains the depth of the human structure, unlike several other models [72, 74, 77] that fail to do so. Further, it visually compared the segmentation maps generated by the SUIM-Net model using the enhanced input from the proposed method and existing UIE methods. The segmentation maps generated with the enhanced image produced by the proposed model exhibited higher class consistency across the entire image.

5.3 Summary

This chapter proposes a multi-degraded representation learning-based multi-domain feature-swapping model for enhancing multi-degraded underwater images. To improve the model's ability in capturing the contextual relationship, first, it leverages cross-attention between spatial and frequency features where frequency information is obtained via the Fast Fourier

Transform. Next, the benefits of integrating prompts into the model are explored to facilitate interrelationships through cross-attention. Experiments conducted on standard underwater datasets demonstrate that the proposed model effectively removes color distortion, haze, blur, and low contrast. Moreover, ablation studies have confirmed that the key components of the proposed model, as well as the carefully combined loss function, are effective in addressing the challenges posed by difficult underwater images.

Moving beyond oceanic conditions, underwater imaging in rivers presents unique challenges due to the presence of clay, silt, sand, and other particles, resulting in blurred and highly non-uniform scenes. In the next contributory chapter, these problems will be considered.

Figure 5.9: *The first, third, fifth, and seventh row shows the enhancement results of existing and the proposed methods. The second row displays the edge detection result of the first row. The fourth row presents the depth map of the third row. The sixth row illustrates the segmentation result of the fifth row. The eighth row features the pie chart of colors of the seventh row.*

"The true laboratory is the mind, where behind illusions we uncover the laws of truth."

~Jagadish Chandra Bose

6

Generating and Enhancing River-water Images

Despite extensive work, the study and analysis of underwater image enhancement algorithms still need to focus on river-based scenarios where the images are muddy due to the clay, silt, sand, and other particles [184]. As of now, no such datasets are available to perform the enhancement task. Additionally, it is almost inconceivable to click a real scene and the complementary ground truth image for different water types at the same time [185].

To address the above issue, first, a muddy water dataset has been generated that is more likely to be a collection of river images. The principal component analysis-based [186] fusion is used between existing underwater images and a few sample muddy water images from river environments to achieve this. Next, a thorough investigation of several SOTA UIE models was conducted using the newly generated dataset of muddy water. Extensive

Figure 6.1: *Examples of degraded underwater images in the ocean (first row), the generated muddy images for the corresponding ocean images (second row), and the available ground truth images (third row).*

experimentations are performed to assess the proposed dataset both quantitatively and qualitatively. In addition, a novel enhancement model was proposed for the muddy water dataset, which is capable of handling degradation present in the images collected from a river environment. To the best of current knowledge, this is the first effort to generate a muddy water dataset and propose a muddy water image enhancement method. Figure 6.1 shows the example of images from different environments along with their ground truths. Major contributions are as follows:

- Generated a muddy water dataset using the principal component analysis (PCA) based-fusion technique, which can be considered an alternative to the images collected from the river environment. This dataset offers a new research direction for the enhancement of degraded images collected from the river environment.

- A complete analysis of several SOTA UIE methods was conducted using the muddy water dataset. The qualitative and quantitative evaluation provides valuable insights into the strengths and limitations of recent UIE methods, which are mainly trained and evaluated on images collected from the ocean environment.

Figure 6.2: *Visual comparison of generated muddy water images with respect to different α values for a few sample images.*

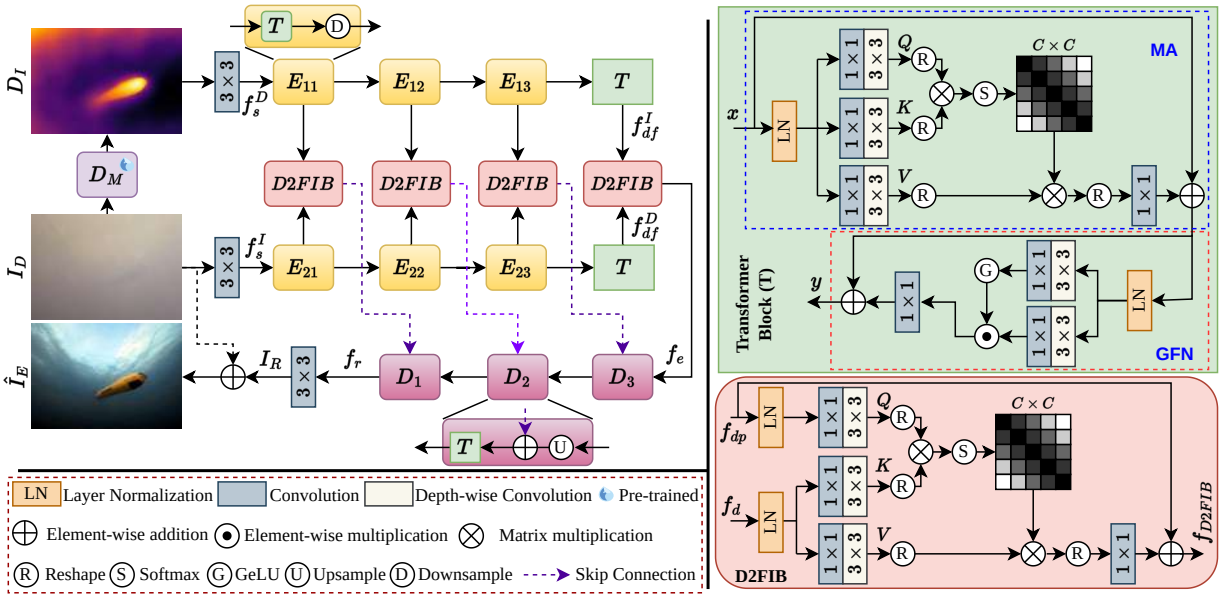


Figure 6.3: *Overview of the proposed model along with all the sub-components.*

- Finally, a novel depth-guided model was proposed to enhance the quality of the muddy water images, demonstrating the generated dataset's generalization and establishing a benchmark for muddy water image enhancement.

6.1 Methodology

This section provides the detailed steps for muddy water dataset generation, followed by a description of the various modules of the proposed model for muddy water image enhance-

ment.

6.1.1 Data Generation

Unlike the ocean environment, the visibility inside the river is very low due to soil-mixed water (muddy water). Due to low visibility and very costly image acquisition equipment, real images of river water are rarely found. Therefore, a new river water dataset was synthetically generated from existing underwater datasets by imposing river water degradation (because of muddy water). A few river water images were collected as style images with different ranges of noise without any detectable objects in the background. Image fusion techniques were utilized to adapt an underwater image collected from the ocean to a river environment. Unlike GAN, the PCA-based method requires a few sample images and does not require substantial training resources and time. Hence, PCA-based image fusion was employed to generate the muddy water dataset. The fusion was performed between content images (existing underwater images) and style images (collected images from the river environment). The PCA-based fusion process is given in an algorithm 2.

As shown in Algorithm 2, there are four major steps in the PCA-based fusion method. (1) Singular value decomposition (SVD) [187] is performed on both content and style images as it computes the principal components directly from the image, avoiding the need for covariance matrix calculations. (2) The top three principal components are selected for capturing the most significant information from the input images while discarding redundant or less important details. (3) The image is fused by combining the selected principal components with the weight factor α. Here, α controls the noise level in the generated muddy water images. (4) Lastly, inverse PCA is carried out to transform the data back into the original image domain. Finally, two muddy water datasets, muddy-UIEB and muddy-LSUI, are generated using the UIEB [30] and LSUI [74] datasets as content and sample muddy images as a style. Multiple α values, i.e., 0.1, 0.2, and 0.3, are considered to maintain the diversity of noise level in the muddy water dataset. Sample-generated images are shown in Figure 6.2 to compare the visual results with respect to different α values for a few sample images.

Algorithm 2 Muddy Water Image Generation

Require: Style image I_s, Content image I_c
Ensure: Generated muddy-water image I_D

Step 1: Perform SVD
$I_c = U_{I_c} \Sigma_{I_c} V_{I_c}^T$
$I_s = U_{I_s} \Sigma_{I_s} V_{I_s}^T$ ▷ U_{I_*} is an orthogonal matrix (left singular vectors), Σ_{I_*} is a diagonal matrix of singular values, $V_{I_*}^T$ is orthogonal matrix (right singular vectors, or principal component directions).

Step 2: Taking first three principal components
$I_s^3 = U_{I_s} \Sigma_{I_s,3} V_{I_s,3}^T$
$I_c^3 = U_{I_c} \Sigma_{I_c,3} V_{I_c,3}^T$ ▷ where $\Sigma_{I_*,3}$ contains the top 3 singular values and $V_{I_*,3}$ contains the corresponding right singular vectors.

Step 3: Combining them with weighted addition
$M = \alpha I_s^3 + (1-\alpha) I_c^3$

Step 4: Inverse PCA Reconstruction
$I_D = U_{I_c} \Sigma_M V_{I_c}^T$ ▷ where Σ_M is the modified diagonal matrix after the weighted addition.
return I_D

6.1.2 Proposed Image Enhancement Model

For the enhancement task, let there be a muddy image dataset having N paired images $DS = \{I_{GT}^k, I_D^k\}$, where I_{GT} and I_D denote the ground truth and degraded images, respectively. The goal is to design a model to enhance the quality of the muddy water image by reducing turbidity. Figure 6.3 shows the overview of the proposed model.

Overall Pipeline. Given a muddy water image I_D, it was fed to a pre-trained Depth-anything model [188] D_M to extract the corresponding depth map D_I. Next, both I_D and D_I are passed through two different convolution layers with a kernel size of 3×3 to obtain the shallow features f_s^I and f_s^D. The shallow features are parallelly fed into two different three-level encoder blocks followed by a transformer block and get the corresponding features f_{df}^I and f_{df}^D. Each encoder block (highlighted by yellow color) contains a transformer block and a down-sampling operation. The extracted features of image I_D and depth map D_I at each level of the encoder are fused through the cross-attention mechanism, namely, depth-degraded feature interaction, to a single encode feature using $D2FIB$ module (highlighted by pink color). The $D2FIB$ module also fuses the final features of image I_D and depth

map D_I, i.e., f_{df}^I and f_{df}^D to a single encoded feature f_e. The D2FIB module learns the correlation between the depth features and degraded features. Next, the encoded feature f_e is passed to a three-level decoder block (highlighted by red color) to get reconstructed features f_r. Each decoder block contains a transformer block and up-sampling operations. To improve the feature reconstruction at each level of the decoder block, the corresponding correlated features from D2FIB are added to the decoder via skip connection. Finally, a 3×3 convolution is applied on f_r to obtain the residual image I_R. Final enhanced image $\hat{I}_E$ is obtained by adding I_R and I_D. The model is optimized using the $\mathcal{L}_1$ loss [189] function:

$$\mathcal{L}_1 = \frac{1}{N} \sum_{i=1}^{N} \|\hat{I}_E^i - I_{GT}^i\|_1 \tag{6.1}$$

6.1.2.1 Transformer Block

A traditional image enhancement vision transformer [141] has been integrated into the model. It comprises two main components: (a) the Multi-Dconv head transposed attention (MA), which enables learning of global relationships across the entire image, tracking connections between distant pixels and selectively focusing on essential regions; and (b) the Gated-Dconv feed-forward network (GFN), which controls the flow of information through the network, allowing only the most informative features to pass forward while suppressing or attenuating less useful ones. This enhances both efficiency and performance. Let x be an intermediate feature input to the transformer, then its output:

$$y = GFN(MA(x)) \tag{6.2}$$

6.1.2.2 Depth-degraded feature interaction Block

Depth image features provide structural and geometric information, such as the distance and shape of objects, which is largely unaffected by visual degradation (e.g., turbidity, blurring, or color distortion). Degraded image features, on the other hand, contain texture, color, and appearance details but are affected by the underwater environment. A Depth-

degraded feature interaction block is proposed, allowing the model to dynamically focus on relevant features from both the depth and degraded images by computing attention scores that indicate the importance of specific regions. This mechanism ensures that the relevant regions in the degraded image are enhanced based on the depth information (e.g., focusing on important objects or areas).

Four D2FIBs with the same architecture are proposed for message passing between depth and degraded features. As evident from Figure 6.3, each D2FIB performs a cross-attention between depth (f_{dp}) and degraded (f_d) features. Since degraded image features contain texture, color, and appearance details, they are kept as key and value features. Message tokens for depth and degraded image features are obtained and passed through a cross-attention (CA) module:

$$f_{D2FIB} = CA(LN(f_{dp}), LN(f_d)) \tag{6.3}$$

6.2 Experiments

6.2.1 Dataset Details

The proposed model is trained and evaluated on two generated muddy water datasets: muddy-UIEB and muddy-LSUI. The muddy-UIEB dataset contains 890 labeled images, split into 800 for training and 90 for testing. The muddy-LSUI dataset consists of 4,279 labeled images, divided into 3,879 for training and 400 for testing.

6.2.2 Implimentation Details

All experiments are conducted using the PyTorch framework on a single Nvidia A100 GPU with a batch size of 5. The model is trained for 400 epochs with an $L1$ loss function, Adam optimizer, and a learning rate of 2e-4. The proposed model is evaluated using the SSIM [113], PSNR [113], and UIQM [117].

6.2.3 Quantitative Result

Table 6.1 shows the results for the muddy-UIEB and muddy-LSUI test sets. For the muddy-UIEB, the proposed method outperformed all existing approaches [72,74,75], with minimum improvements of 3.42% in SSIM and 5.04% in PSNR. For the muddy-LSUI, the minimum improvements are 2.55% in SSIM and 2.54% in PSNR. These results demonstrate the proposed model's effectiveness in maintaining both high fidelity and perceptual quality.

Table 6.1: *Quantitative result of proposed model and SOTA models.*

Model	SSIM	PSNR	UIQM
muddy-UIEB			
SGUIE-Net [75]	0.497	15.595	**3.396**
DeepWave-Net [72]	0.451	14.033	3.184
Ushape [74]	0.440	16.034	2.964
Proposed	**0.514**	**16.842**	3.199
muddy-LSUI			
SGUIE-Net [75]	0.695	20.906	**3.309**
DeepWave-Net [72]	0.674	18.617	3.288
Ushape [74]	0.706	23.064	3.184
Proposed	**0.724**	**23.649**	3.126

6.2.4 Qualitative Result

The proposed model's effectiveness was assessed by visually comparing its enhanced images with those from state-of-the-art models, as shown in Figure 6.4. The model achieves significantly better image quality, likely due to the use of depth maps to guide feature extraction and capture fine details of objects in degraded images. The SGUIE-Net and Ushape models also perform well and are comparable to the ground truth, using semantic maps and different color spaces for enhancement, respectively. These findings highlight the importance of incorporating complementary information, with depth maps being the most effective for enhancing muddy water images.

Additionally, to validate the model on real-world muddy water images, visual results of sample images collected from the river environment are shown in Figure 6.5. As a first contribution in this domain, the proposed method produces visually appealing results for

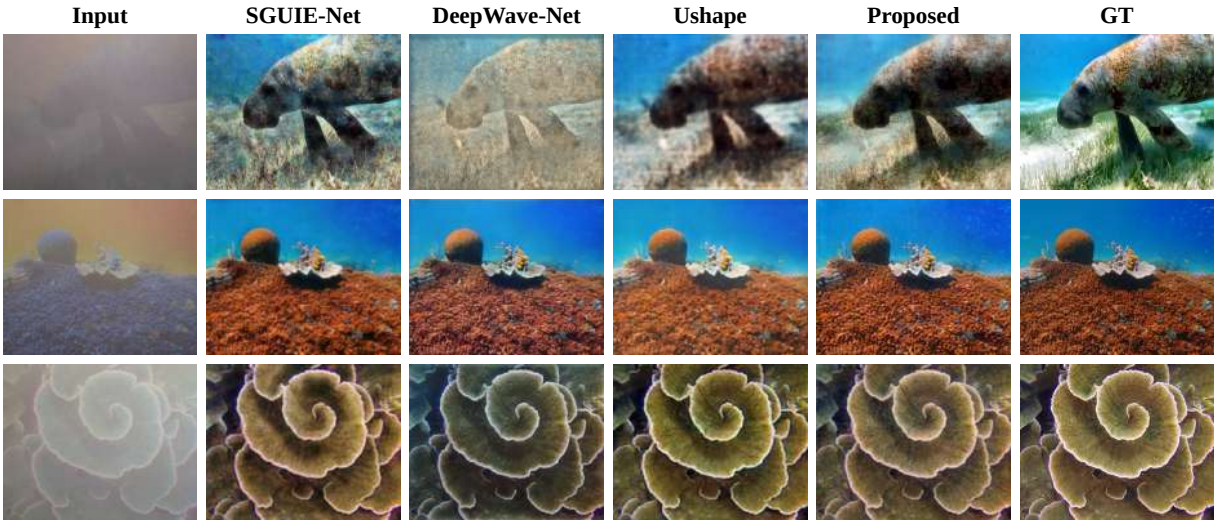


Figure 6.4: *Visual comparison of the proposed and SOTA methods.*

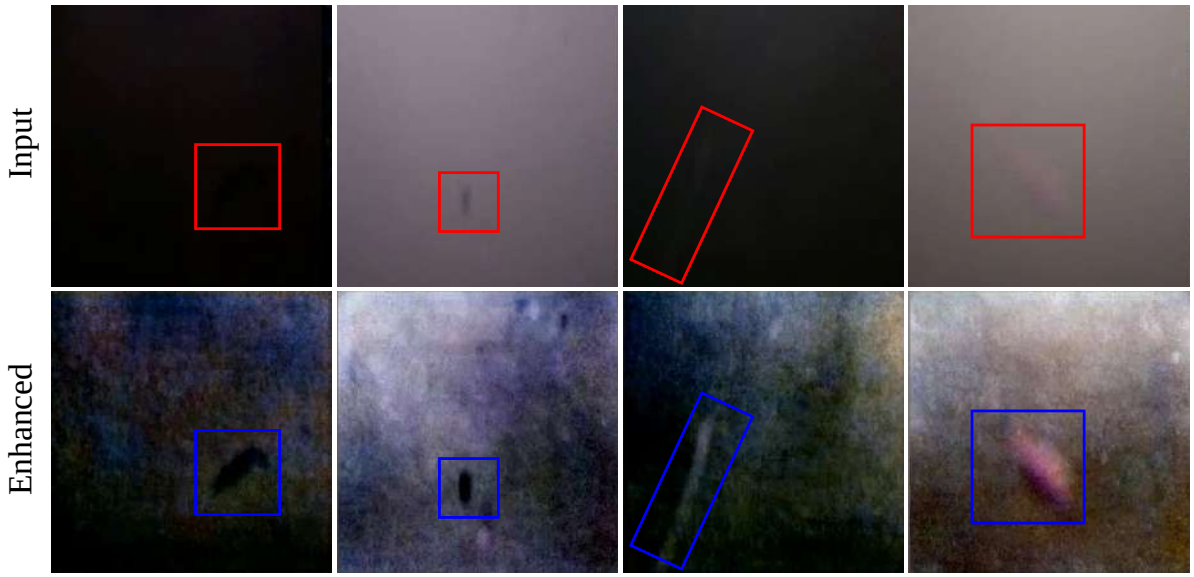


Figure 6.5: *Validation on real-world muddy water images.*

highly degraded river images.

6.2.5 Ablation Study

Different ablation studies were executed to validate the effectiveness of the proposed model. These are:

- $\mathcal{AB}\mathbf{1}$: The depth map was removed from the proposed model, and only the degraded image is used for enhancement.

- $\mathcal{AB}$**2**: An addition was chosen inside D2FIB.
- $\mathcal{AB}$**3**: The skip connections were removed.

Table 6.2 provides a comprehensive justification of the utility of different settings, further confirmed by the clear visual comparisons illustrated in Figure 7.6.

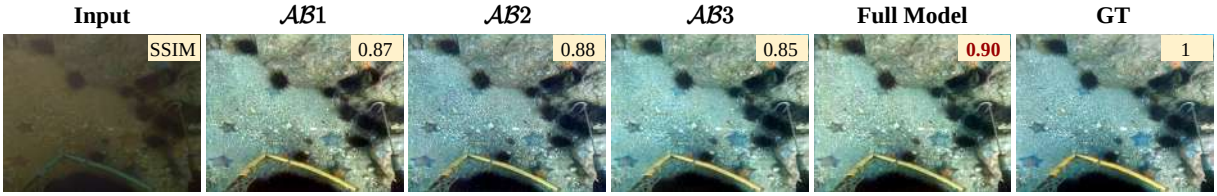


Figure 6.6: *Visual comparison with different settings in proposed model.*

Table 6.2: *Quantitative comparison of different settings of the proposed model.*

Model	SSIM	PSNR	UIQM
muddy-UIEB			
$\mathcal{AB}$**1**	0.460	16.171	**3.335**
$\mathcal{AB}$**2**	0.415	15.066	3.329
$\mathcal{AB}$**3**	0.444	15.559	3.142
Full Model	**0.514**	**16.842**	3.199

6.3 Summary

This chapter proposed a framework called 'River-GEM' that generates and enhances muddy water images. First, this work evaluates the existing UIE models using two generated datasets, i.e., muddy-UIEB and muddy-LSUI, to understand their limitations and establish a baseline. Based on the observation, a depth-guided enhancement model is proposed to improve the quality of the muddy water images and establish a new benchmark for generated muddy water datasets. The quantitative and qualitative results from extensive experimentation and ablation show the superiority of the proposed model over existing SOTA schemes. The proposed work provides a new direction for research aimed at enhancing the muddy water images collected from the river environment, which exhibits a different set of degradation compared to images collected from the ocean environment.

So far in this dissertation, the enhancement of ocean-water and river-water images has been addressed using a variety of deep learning-based models. Although enhancement improves color balance, contrast, and visibility, the images often still lack important details. This happens because low-resolution underwater sensors cannot capture fine textures, sharp edges, and precise details during image acquisition. As a result, important visual information may still look blurry or unclear, even after enhancement. With this motivation, the next chapter focuses on underwater image super-resolution, which aims to reconstruct high-resolution images from low-resolution underwater inputs.

The work presented in this chapter has been published at the ICASSP 2025 conference.

"The important thing is not to stop questioning. Curiosity has its own reason for existing."

~Albert Einstein

7

Dual-domain Attention Network for Underwater Image Super-resolution

Underwater imaging is a challenging and critical field of study with applications in marine biology, underwater exploration, and surveillance [190]. Nonetheless, underwater images frequently suffer from degradation caused by the absorption of light in water, leading to poor visibility, color distortion, and blurring [30]. These degradations can significantly impact the quality and resolution of underwater images, making them challenging to analyze and interpret. SR is a useful method to improve an image's resolution, resulting in a higher level of detail and clarity, either by extrapolating or interpolating the existing pixels or by using additional information from similar images [191]. In this final contributory chapter, the two

works for UISR are discussed.

Figure 7.1: *Visual demonstration of 3X super-resolve example on UFO-120 dataset.*

7.1 Attention-based Spatial-Frequency Information Network for Underwater Image Super-resolution

Traditional techniques like bicubic and nearest-neighbor interpolation rely on the consistency of neighboring pixels to reconstruct unknown pixels [105]. The main limitation of these models lies in their inability to produce satisfactory reconstruction due to limited feature representational capabilities. Among existing studies, the learning-based techniques that acquire the ability to map between LR and HR image space have reached state-of-the-art performance. GAN-based residual learning [108], attention-based multiple cross-paths convolution [110], color features-based [30], and wavelet transformation-based deep network [106] have been used in the UISR task. In [109], the authors suggested a loss function considering image sharpness, color casts, and overall contrast enhancement for UISR. Recently, Sharma et al. [72] proposed a framework for UISR using wavelength-based attributes. Despite significant improvements in UISR methods, they still suffer from visual artifacts in the SR images, such as color distortion, lack of texture representation, and loss of detail. Also, it is observed that the high-frequency image regions are more complex to super-resolve than the low-frequency ones. Most existing UISR methods operate within the spatial domain, focusing on enhancing the resolution of LR pixels to get an HR image. These techniques neglected the use of the frequency domain, which could offer a more straightforward approach to recovering missing high-frequency information, as shown in some restoration tasks [192, 193].

Keeping the above limitations in mind, this work proposes an architecture that amalgamates DWT and effective utilization of attention modules to enhance the performance of UISR. The emphasis lies not solely on the super-resolution performance but also on ensuring the content and structural consistency between the predicted and ground-truth images. Figure 7.1 demonstrates the visual result of the UISR technique, while the proposed approach delivers better visuals with enhanced color correction and finer detail. The primary contributions of this work can be outlined as follows: **1)** Both frequency and spatial information are used for the proposed restoration scheme. **2)** Different wavelet sub-bands are analyzed separately, enabling a more accurate restoration of high-frequency details and improving the super-resolve performance. **3)** An attention block is used to suppress irrelevant features and extract useful refined features that increase the representation power of the model. **4)** The L1 loss is replaced with Color channel-specific L1 loss to super-resolve underwater images considering the different behavior of various color channels in underwater images.

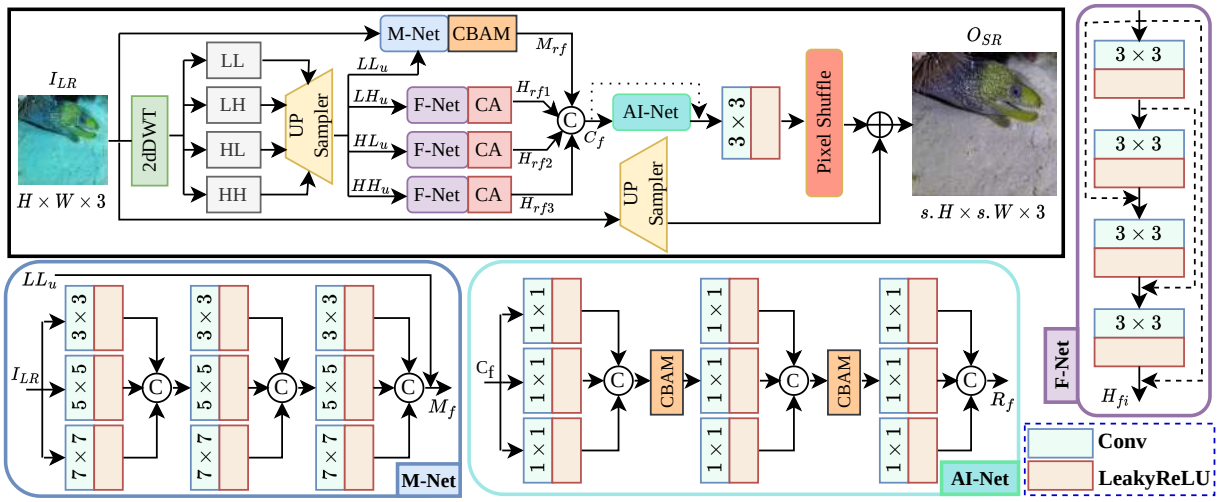


Figure 7.2: *Overview of the proposed model for UISR. It accepts a degraded LR image as input and produces an SR image that is visually and spatially enhanced. s.H/s.W denotes the image height/width scaled by factor s. CA indicates the channel attention.*

7.1.1 Proposed Model

Figure 7.2 shows the proposed architecture that aims to generate a SR image from the degraded LR image. In this work, to utilize the frequency domain information along with the

spatial domain, the Haar Wavelet Transformation (HWT) [194] is applied to the degraded LR image, decomposing it into LL, LH, HL, and HH sub-bands. The LL sub-band represents the background details in the image, while the LH sub-band illustrates variations along the vertical axis, the HL sub-band captures variations along the horizontal axis, and HH contains diagonal information of the image. In general, the dyadic partitioning of the sub-band LL is utilized for multi-resolution analysis. Since LH, HL, and HH sub-bands retain high-frequency details, these three sub-bands are more informative for super-resolving high-frequency information. Additional important background details for UISR are retained from the original image, which is also used along with wavelet sub-bands.

Let $I_{LR} \in \mathbb{R}^{H \times W \times 3}$ be the degraded LR image given as input to the proposed model. First, the LR image is decomposed into LL, LH, HL and HH sub-bands, using HWT where every sub-band $\in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 1}$. The sub-bands are expanded spatially by a factor of two so that $LL_u = U_b(LL)$, $LH_u = U_b(LH) \in \mathbb{R}^{H \times W \times 1}$, $HL_u = U_b(HL) \in \mathbb{R}^{H \times W \times 1}$, and $HH_u = U_b(HH) \in \mathbb{R}^{H \times W \times 1}$ where U_b represents the Bicubic interpolation. Then the I_{LR} and LL_u are passed to **M-Net** and processed in the spatial domain. LH, HL, and HH are passed into **F-Net** and processed in the frequency domain. The proposed M-Net takes I_{LR} and LL_u as input and uses three different sizes of receptive fields to capture multi-contextual features ($M_f = MNet(I_{LR}, LL_u)$), where $MNet$ denotes the M-Net block, consisting of three sequential layers; each contains three parallel convolutional layers with distinct kernel sizes, 3×3, 5×5, and $\times 7$, respectively. This module captures diverse contextual information by employing varying kernel sizes. The major rationale behind this setup is to use a large kernel to model global context while a small kernel to capture the local ones. This is particularly advantageous when dealing with images with complex and diverse structures. Further, the multi-contextual spatial features are refined using a CBAM [1] and get $M_{rf} = CBAM(M_f)$. CBAM takes an intermediate multi-contextual feature map as input and determines '**what**' aspects of the features significantly represent the desired information and concentrates on '**where**' is an informative part.

Each LH_u, HL_u, and HH_u is passed through F-Net, which utilizes the frequency in-

formation and helps to super-resolve the high-frequency regions by capturing the high-frequency features ($H_{fi}, i = 1, 2, 3$). Then, the irrelevant information is suppressed and focus on **what** is meaningful given H_{fi} by using channel attention and get H_{rfi}, where $H_{rf1} = CA(H_{f1}), H_{f1} = FNet(LH_u)$, $H_{rf2} = CA(H_{f2}), H_{f2} = FNet(HL_u)$, and $H_{rf3} = CA(H_{f3}), H_{f3} = FNet(HH_u)$. As these sub-bands contain high-frequency features, retaining

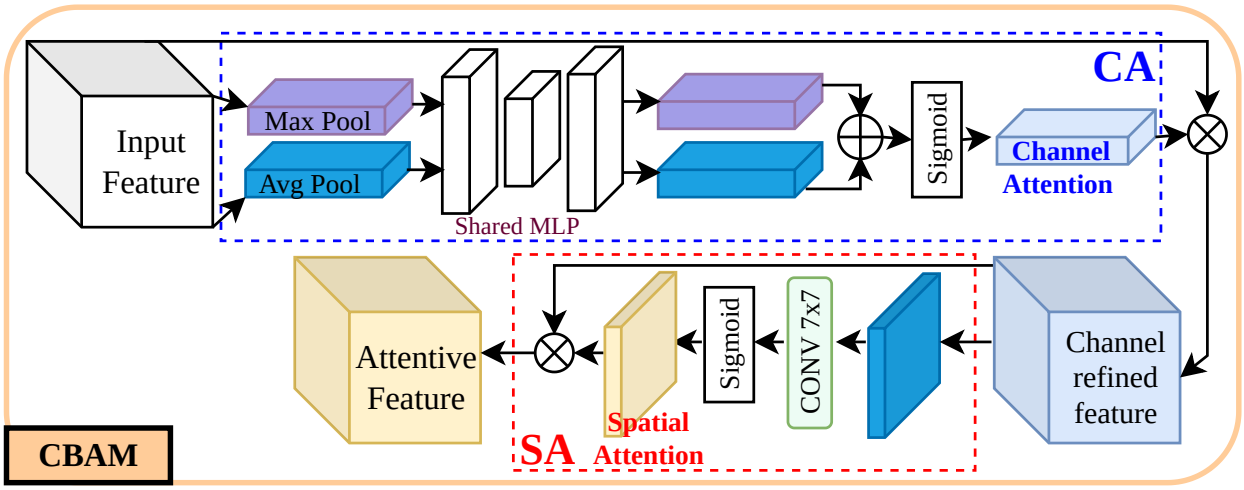


Figure 7.3: *Network architecture of CBAM [1].*

ing them when passing them through the network is essential to avoid losing the necessary details. The F-Net architecture employs multiple skip connections to meet this requirement by retaining the original information from the input feature. By doing so, the network can preserve high-frequency information and utilize it to generate more accurate and detailed results. These intermediate feature maps are concatenated and fed into **AI-Net** to refine further, where $R_f = AINet(C_f = M_{rf} \odot H_{rf1} \odot H_{rf2} \odot H_{rf3})$. AI-Net used multiple layers of 1x1 convolution and CBAM blocks. Use of 1x1 convolution layers enables maximum information retention while reducing the feature dimensions, which is important for generating good SR images. This approach ensures the network's final output contains the relevant features required for generating high-quality SR images. The refined features pass through a convolution and a pixel-shuffle [195] layer to enhance the spatial resolution of the underwater image. Finally, the SR image is obtained by: $O_{SR} = PS((c^{3\times3}(l(R_f \odot C_f)))) + U_b(I_{LR})$.

Methods	PSNR			SSIM			UIQM		
	2x	3x	4x	2x	3x	4x	2x	3x	4x
SRResNet [196]	25.23 ± 4.1	23.85 ± 2.8	19.13 ± 2.4	0.74 ± 0.08	0.68 ± 0.07	0.56 ± 0.05	2.42 ± 0.37	2.18 ± 0.26	2.09 ± 0.30
SRCNN [197]	24.75 ± 3.7	22.22 ± 3.9	19.05 ±2.3	0.72 ±0.07	0.65 ±0.09	0.56 ± 0.12	2.39 ± 0.35	2.24 ±0.17	2.02 ± 0.47
SRGAN [196]	26.11 ± 3.9	23.87 ± 4.2	21.08 ± 2.3	0.75 ± 0.06	0.70 ± 0.05	0.58 ± 0.09	2.44 ± 0.28	2.39 ± 0.25	2.26 ± 0.17
SRDRM [108]	24.62 ± 2.8	-	23.15 ± 2.9	0.72 ± 0.17	-	0.67 ± 0.19	2.59 ± 0.64	-	2.56 ± 0.63
SRDRM-GAN [108]	24.61 ± 2.8	-	23.26 ± 2.9	0.72 ± 0.17	-	0.67 ± 0.19	2.59 ± 0.64	-	2.57 ± 0.63
Deep SESR [109]	25.70 ± 3.2	26.86 ± 4.1	24.75 ± 2.8	0.78 ± 0.08	0.75 ± 0.06	0.66 ± 0.05	3.15 ± 0.48	2.87 ± 0.39	2.55 ± 0.35
DWNet [72]	25.71 ± 3.0	25.23 ± 2.7	25.08 ± 2.9	0.80 ± 0.07	0.79 ± 0.07	0.74 ± 0.07	2.99 ± 0.57	2.96 ± 0.60	2.97 ± 0.59
Proposed	26.33 ± 2.7	25.65 ± 2.5	25.31 ± 2.8	0.83 ± 0.07	0.82 ± 0.07	0.81 ± 0.07	2.93 ± 0.56	2.88 ± 0.58	2.86± 0.55

Table 7.1: *Comparison between proposed model with the State-of-the-Art methods on UFO-120 dataset. Red, blue, and green indicate best, second-best, and third-best values, respectively.*

7.1.1.1 Loss Function

Most of the recent UISR schemes use L1 loss, but it is observed that color channel-wise L1 loss may be a better option for underwater images as the behavior of these images are different for different color channels. In this work, color channel-wise L1 loss is used to preserve the essential information for the underwater images. Let O_{SR} be the super-resolved image estimated by the proposed model and I_{SR} be the ground truth HR image. The color channel-wise L_1 loss function can be written as:

$$L_{1_{channel}} = \lambda_r.L_{1_r} + \lambda_g.L_{1_g} + \lambda_b.L_{1_b}. \tag{7.1}$$

where, $L_{1_r} = \frac{1}{a}\sum_{j=1}^{a} \|((O_{SR})_j)_r - ((I_{SR})_j)_r\|_1$, $L_{1_g} = \frac{1}{a}\sum_{j=1}^{a} \|((O_{SR})_j)_g - ((I_{SR})_j)_g\|_1$, and $L_{1_b} = \frac{1}{a}\sum_{j=1}^{a} \|((O_{SR})_j)_b - ((I_{SR})_j)_b\|_1$. Experimentally this work sets $\lambda_r = 1$, $\lambda_g = 1.5$, and $\lambda_b = 2$.

Since L_1 does not align closely with how humans perceive image quality, the **perceptual loss** [122] is employed to preserve high-frequency details of the image. To achieve this, a VGG-16 model ($V(.)$) is used and pre-trained on ImageNet to extract features specifically from the $relu2\_2$ layer. The perceptual loss can be represented as:

$$L_{VGG} = \frac{1}{a}\sum_{j=1}^{a} \left\|V((O_{SR})_j) - V((I_{SR})_j)\right\|_2 \tag{7.2}$$

Apart from $L_{1_{channel}}$ and L_{VGG}, also **SSIM loss** [114] is used that quantifies the similarity between two images by considering structural similarity. The SSIM loss function can be

written as:

$$L_{SSIM} = \frac{1}{a}\sum_{j=1}^{a}(1 - SSIM((O_{SR})_j, (I_{SR})_j)) \tag{7.3}$$

Lastly, the **DoG loss** is employed to preserve the quality and enhance the sharpness of edges in an SR image, as the edges become blurry when an LR image is super-resolved. Five Gaussian kernels of size 7x7 are generated with σ, $k\sigma$, $k^2\sigma$, $k^3\sigma$, $k^4\sigma$ as their standard deviation. These kernels are convolved over the predicted SR and Ground Truth images, and then MSE loss is calculated between them to find the DoG loss value. σ = 1.6 and k = $\sqrt{2}$ were taken, like [193]. The overall loss function that needs to be optimized can be written as:

$$L_T = L_{1_{channel}} + \lambda_v.L_{VGG} + \lambda_s.L_{SSIM} + L_{DoG} \tag{7.4}$$

where emperically $\lambda_v = 0.02$, $\lambda_s = 0.5$ are set.

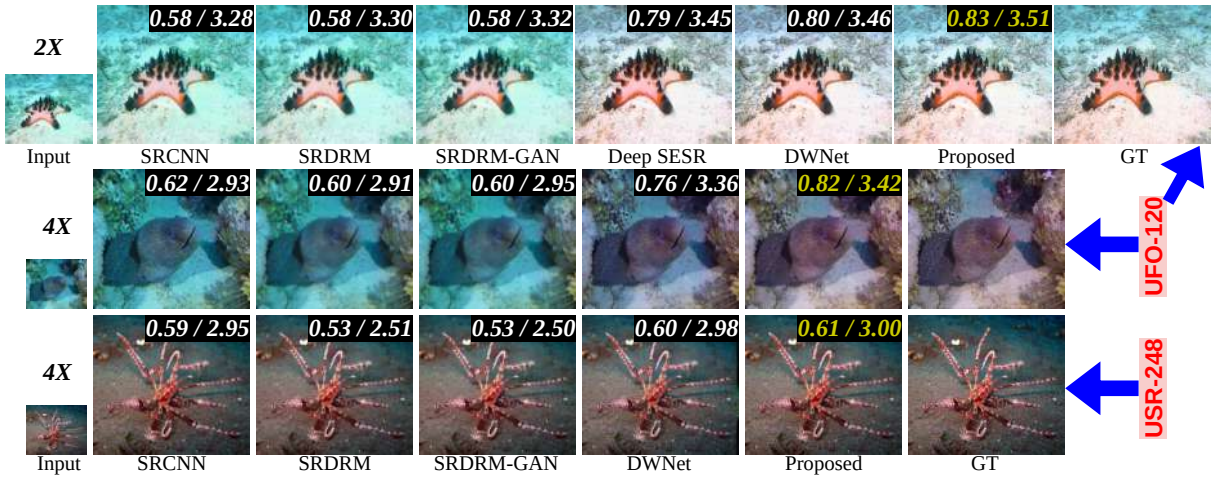


Figure 7.4: *Visual comparison of the proposed model with existing works on UISR. SSIM/UIQM score is marked on the corner.*

7.1.2 Experiments

To evaluate the efficacy of the proposed model, experiments were conducted on two UISR benchmark datasets: UFO-120 [109] and USR-248 [108] and performance was compared

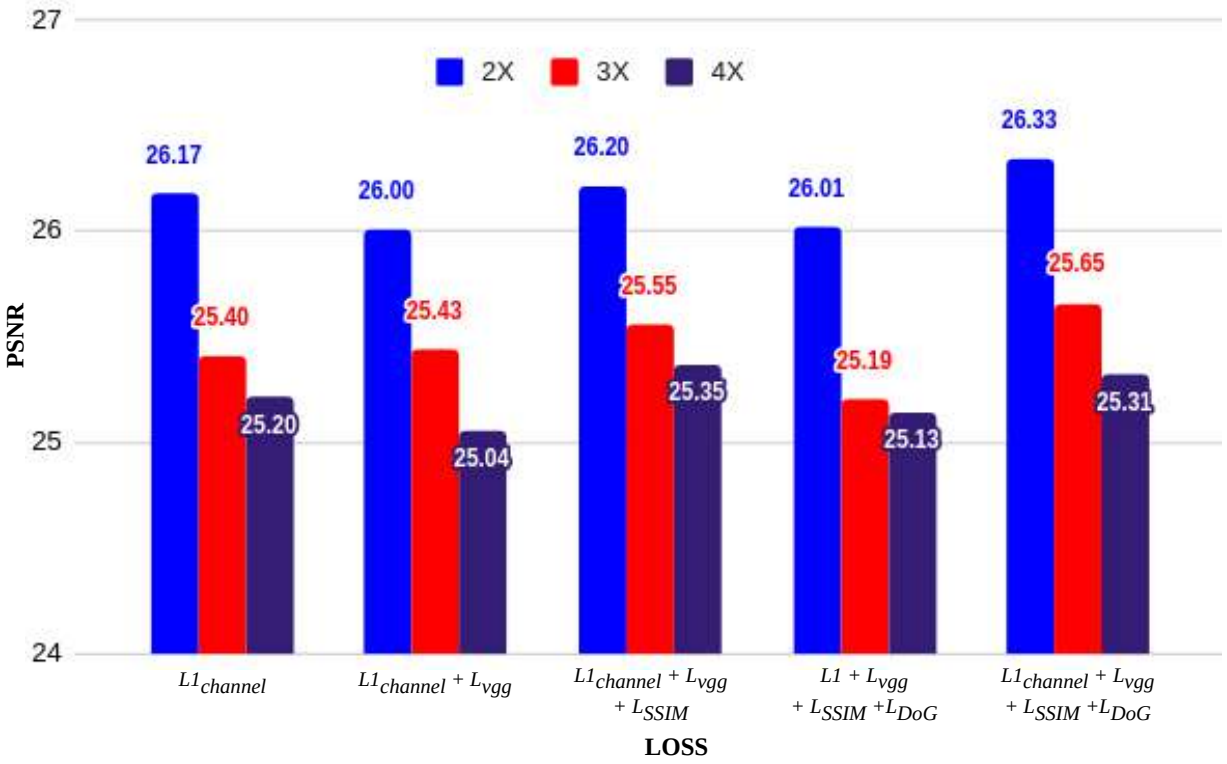


Figure 7.5: *Scale-wise effect of loss functions on UFO-120 dataset.*

with eight state-of-the-art methods. The evaluation metrics used were PSNR, SSIM, and the UIQM [117]. The model is trained using the Pytroch framework with a batch size of 5 and a learning rate of 2e-4 on the Nvidia A100 GPU. In this model, batch normalization layers have not been used to reduce memory consumption, alleviate processing overhead, and improve accuracy.

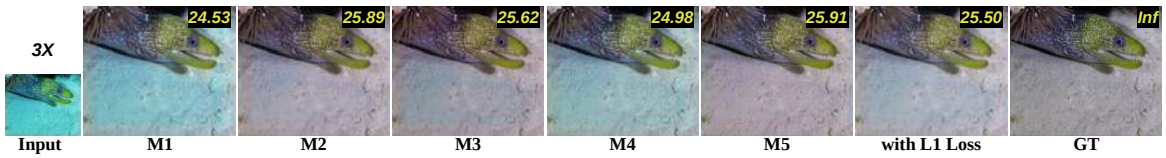


Figure 7.6: *Visual demonstration of different components of the proposed model on UFO-120 (3x). PSNR is marked on the corner.*

Table 7.1 shows that the method improved 0.22 dB and 0.23 dB on 2x and 4x of the UFO-120 dataset. Furthermore, the proposed model yields improved SSIM values, with an increase of 3.75% for 2x, 3.79% for 3x, and 9.45% for 4x scale on the UFO-120 dataset. This indicates that the method effectively recovers luminance and contrast while maintaining

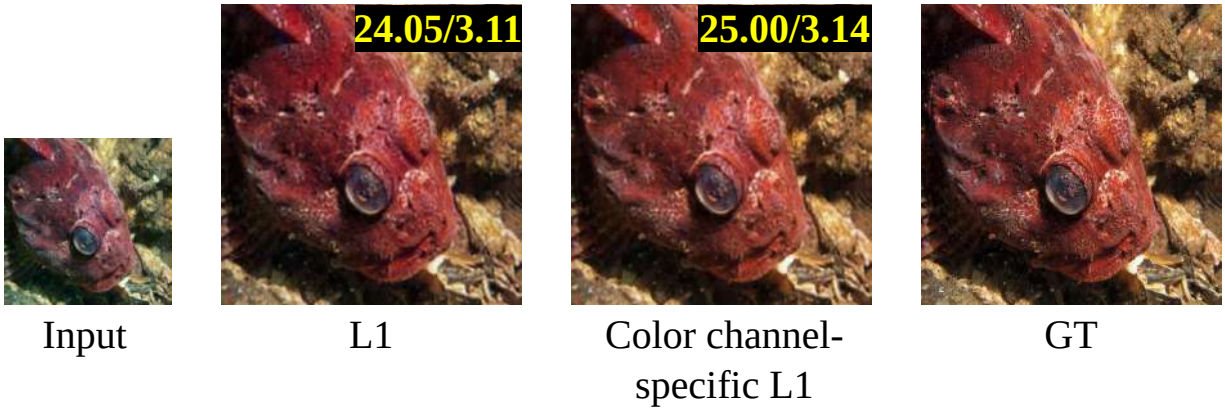


Figure 7.7: *Visual comparison of L1 and color channel-specific L1 loss on UFO-120 dataset. PSNR/UIQM is marked on the corner.*

structural information and fine details. Similarly, Table 7.2 shows that the proposed model surpassed existing methods on the USR-248 dataset. Notably, in the case of UIQM, the proposed model improved over the state-of-the-art by 0.38% for 4x and 6.66% for 8x scale. The qualitative evaluation of UFO-120 and USR-248 datasets is shown in Figure 7.4. Compared with existing methods, the reconstructed SR image of the proposed method achieved outstanding performance and provided more precise textures and richer details.

Ablation studies were conducted on the following aspects: **M1**: Removed the CBAM block from the proposed model, **M2**: Increased the number of AI-Net blocks to 2, **M3**: Set a fixed-size kernel instead of three different sizes of kernels in M-Net, **M4**: Used only one convolution layer in M-Net and AI-Net instead of three parallel convolution layers, and **M5**: Full model. Table 7.5 and Figure 7.6 present the quantitative and qualitative results of different settings, allowing a comprehensive understanding of the impact of various modules within the proposed model. Figure 7.5 illustrates the effect of different loss functions. Visually, the advantage of color channel-specific L1 loss over L1 loss is exhibited in Figure 7.7, which boosts performance and improves color consistency.

Methods	PSNR	SSIM	UIQM
	2x/4x/8x	2x/4x/8x	2x/4x/8x
SRResNet [196]	25.98/24.15/19.26	.72/.66/.55	—/—/—
SRCNN [197]	26.81/23.38/19.97	.76/.67/.57	2.74/2.38/2.01
SRGAN [196]	28.05/24.76/20.14	.78/.69/.60	2.74/2.42/2.10
ESRGAN [191]	26.66/23.79/19.75	.75/.66/.58	2.70/2.38/2.05
SRDRM [108]	28.36/24.64/21.20	.80/.68/.60	2.78/2.46/2.18
SRDRM-GAN	28.55/24.62/20.25	.81/.69/.61	2.77/2.48/2.17
AMPCNet [110]	29.54/25.90/23.83	.80/.66/.62	2.77/2.58/2.25
DWNet [72]	27.86/25.93/—	.76/.68/—	2.71/2.58/—
Proposed	29.70/26.39/23.50	.81/.70/.56	2.76/2.59/2.40

Table 7.2: *Comparison between proposed model with the State-of-the-Art methods on USR-248 dataset.*

Model	PSNR	SSIM	UIQM
M1	25.12 ± 3.1	.80 ± .07	2.96 ± .55
M2	26.12 ± 2.9	.81 ± .07	2.87 ± .64
M3	25.87 ± 2.7	.80 ± .07	2.92 ± .61
M4	25.54 ± 2.7	.79 ± .07	2.92 ± .61
M5	26.33 ± 2.7	.83 ± .07	2.93 ± .56

Table 7.3: *Effect of different components of the proposed model on UFO-120 dataset with scaling factor 2. Red indicates best value*

7.2 Efficient-USR: Prompt Guided Dual-domain Feature Information for Efficient Underwater Image Super-resolution

Despite significant improvements in UISR methods, they still suffer from visual artifacts in the super-resolved images, such as color distortion, lack of texture representation, and loss of detail. It is important to consider features at different scales to better capture and correct these distortions. Also, most previous studies have only analyzed images in the spatial domain for UISR, ignoring the fact that the frequency domain information can help super-resolve the complex high-frequency image regions. Further, these models are computationally expensive and demand a significant amount of memory, posing challenges in resource-constrained or real-time environments.

Considering the above observations, this work proposes a lightweight multi-scale architecture that leverages spatial and frequency domain information to improve the performance

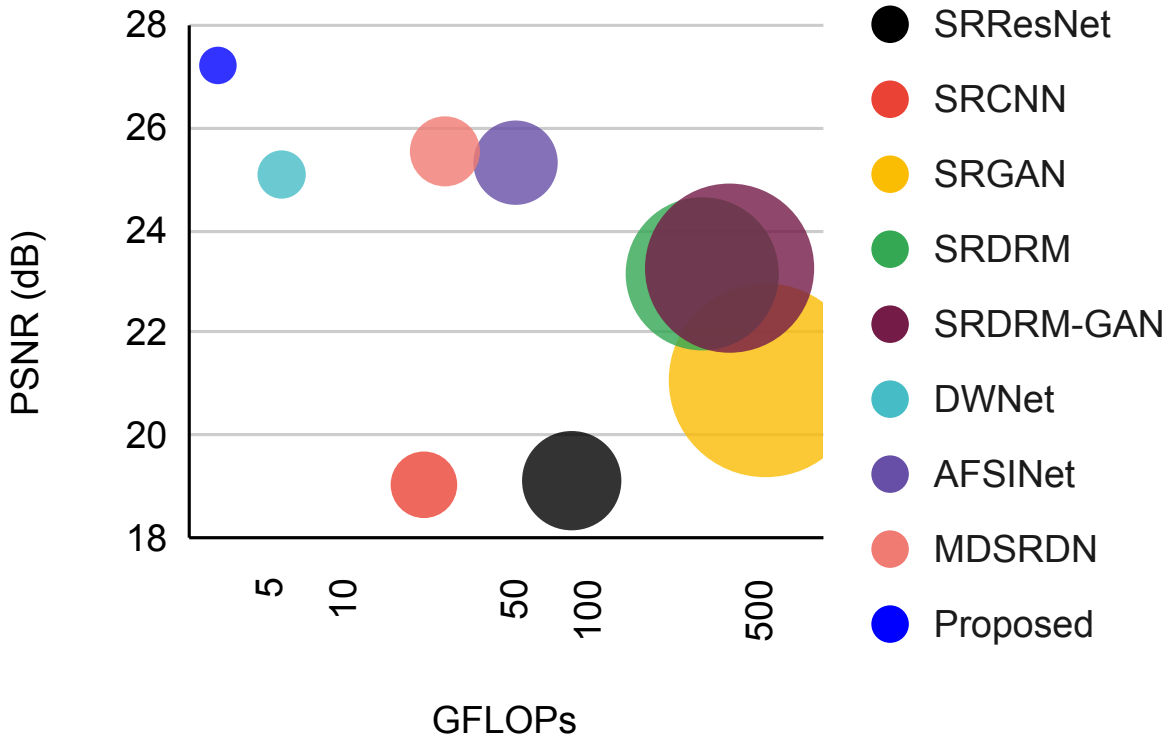


Figure 7.8: *Trade-off between PSNR vs GFLOPs on UFO-120 ×4 dataset. The results show the superiority of EFFICIENT-USR among existing methods.*

of UISR. In contrast, 2D FFT is used for frequency domain analysis to enhance or suppress specific features, such as reducing noise in low frequencies and enhancing details in high frequencies. More recently, image restoration strategies using prompt learning [180, 198, 199] focus on deriving visual prompts from training datasets to optimize the restoration process. To this end, prompt guided cross-attention is utilized to integrate the multiple scale features. Figure 7.8 illustrates the trade-off between PSNR and GFLOPs on the UFO-120 ×4 dataset, while Efficient-USR demonstrates superiority over the SOTA. Contributions include:

- The proposed Efficient-USR model efficiently harnesses the power of dual-domain (spatial and frequency) features to achieve better performance.
- A prompt-guided cross-attention block is incorporated to effectively encode degraded visual information at multiple scales.

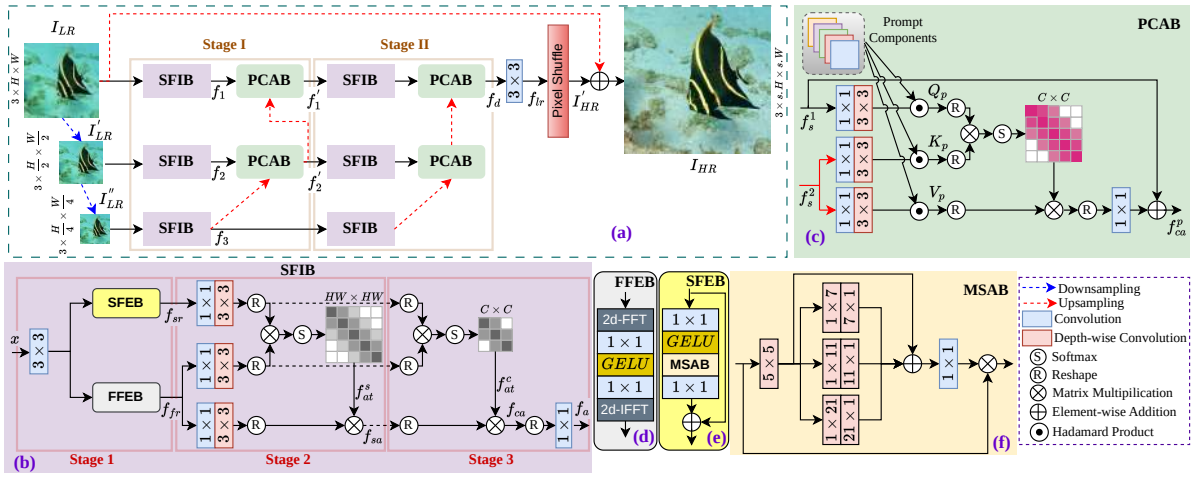


Figure 7.9: *Overview of the proposed Efficient-USR along with all the sub-components.*

7.2.1 Proposed Model

The primary aim is to integrate insights from both the frequency and spatial domain information to reveal fine details and patterns in LR underwater images. Several key designs are incorporated to super-resolve a LR underwater image in the proposed network. First, the overall pipeline of the proposed model is briefly described as depicted in Figure 7.9. Then, two components of the model are introduced: the Spatial-frequency interaction block (SFIB) in Sec. 7.2.1.1 and the prompt-guided cross-attention block (PCAB) in Sec. 7.2.1.2.

Overall Pipeline. Given a degraded LR image ($I_{LR} \in \mathbb{R}^{H\times W\times 3}$), the process begins by down-sampling it twice to obtain two downscaled LR images: $I'_{LR} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$ and $I''_{LR} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3}$. Next, in stage I, these three different scaled images are passed through three SFIBs for spatial-frequency mutual learning, resulting in refined features $f_1 \in \mathbb{R}^{H\times W\times c}$, $f_2 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times c}$, and $f_3 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times c}$ at three different scales. Two prompt-guided cross-attention mechanisms are then performed for message passing between these multi-scale features to enhance the model's ability to understand context: (a) one between up-scaled f_3 and f_2, resulting in f'_2, and (b) another between up-scaled f'_2 and f_1, resulting in f'_1.

In stage II, the features f'_1, f'_2, and f_3 are fed into the next phase to obtain deep features $f_d \in \mathbb{R}^{H\times W\times c}$. A convolution layer is then employed with the output channels set to $3s^2$, where s denotes the scale factor by which the spatial resolution is to be enhanced. Finally, a PixelShuffle (PS) layer generates the high-resolution image $I'_{HR} \in \mathbb{R}^{s.H\times s.W\times 3}$ by taking the

low-resolution feature maps $f_{lr} \in \mathbb{R}^{H \times W \times 3s^2}$ as input. At the end, the upsampled degraded image is added to the I'_{HR}, resulting in the reconstructed HR image ($I_{HR} = I'_{HR} + \mathcal{U}_s(I_{LR})$, where $\mathcal{U}_s$ represents the bi-cubic up-sampling). The proposed Efficient-USR is optimized using the $\mathcal{L}_1$ loss:

$$\mathcal{L}_1(G_{HR}, I_{HR}) = \frac{1}{N} \sum_{j=1}^{N} \|(I_{HR})^j - (G_{HR})^j\|_1, \tag{7.5}$$

where G_{HR} indicates the ground-truth image.

7.2.1.1 Spatial-Frequency Interaction Block (SFIB)

Transformers are proficient at capturing long-range dependencies by calculating attention between query and key vectors. However, due to the blur, color, and contrast distortions in degraded underwater images, focusing attention only on the spatial domain can miss global details, causing unwanted artifacts. In contrast, SFIB is proposed, which processes queries in the spatial domain and keys and values in the frequency domain for a more detailed attention map. Additionally, transformers have shown limitations in capturing local context. In SFIB, attention is calculated in both channel and spatial dimensions to overcome the problem of local context.

In the SFIB framework, the final attentive feature map is generated through three stages. To formally define its principle, let x be an image or intermediate feature map; in stage 1, SFIB passes it through convolution with kernel size 3×3 ($c^{3\times3}$) to get the shallow features f_s. Next, the feature maps f_s are fed into the spatial feature extractor block (SFEB) to extract the spatially refined features f_{sr} and frequency feature extractor block (FFEB) to extract the refined frequency features f_{fr} in parallel. It can be written as

$$f_{sr} = SFEB(c^{3\times3}(x)), \quad f_{fr} = FFEB(c^{3\times3}(x)) \tag{7.6}$$

In stage 2, the spatially attentive (local context) feature map of $\mathbb{R}^{HW \times HW}$ in the spatial dimension is calculated via cross-attention between spatial and frequency domain features. It can be computed as

$$f^s_{at} = CA_s(\mathcal{R}(f_{sr}), \mathcal{R}(f_{fr})), \quad f_{sa} = \mathcal{R}(f_{fr}).f^s_{at} \tag{7.7}$$

where $\mathcal{R}$ denotes the reshape operation, CA_s denotes the cross-attention in the spatial dimension. In stage 3, similar to stage 2, the channel-wise attentive (global context) feature map of $\mathbb{R}^{C\times C}$ is calculated in the channel dimension, as shown in Figure 7.9(b). Lastly, a 1×1 convolution layer ($c^{1\times1}$) is employed to obtain the final feature map f_a. The expression can be formulated as follows:

$$f_{at}^{c} = CA_c(\mathcal{R}(f_{sr}), \mathcal{R}(f_{fr}), \quad f_a = c^{1\times1}(\mathcal{R}(\mathcal{R}(f_{sa}).f_{at}^{c})) \tag{7.8}$$

Spatial Feature Extractor Block (SFEB). For the spatial feature extractor, the framework from [200] has been partially utilized. As depicted in Figure 7.9(e), SFEB processes the input features through a 1×1 convolution, GELU operation, multi-scale attention block (MSAB), and 1×1 convolution operation. MSAB does not use self-attention and instead em-

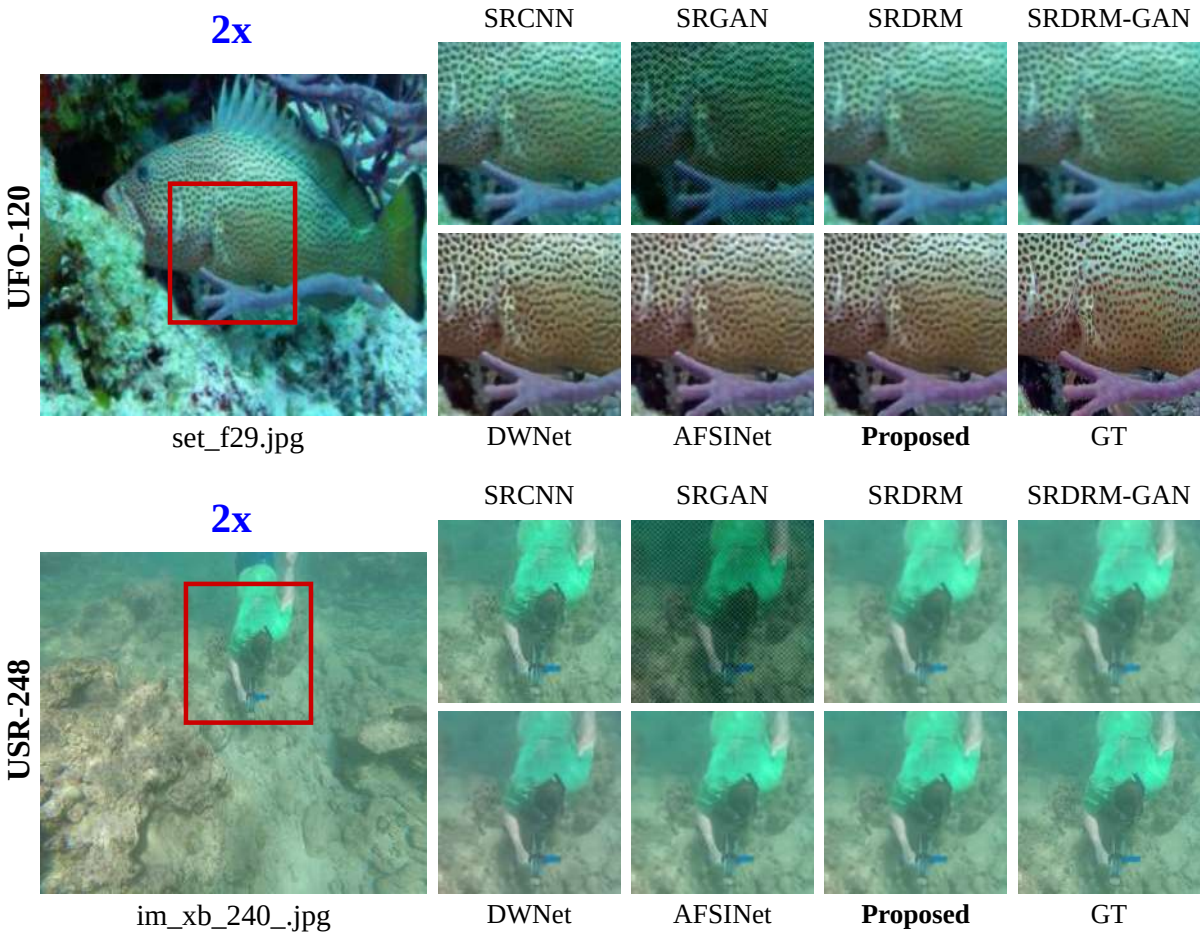


Figure 7.10: *Visual comparison with SOTA methods on UFO-120 (×2) and USR-248 (×2) datasets.*

ploys depth-wise convolution for local context, three parallel depth-wise strip convolutions for multi-scale information, and a 1×1 convolution to represent the associations between various channels. Let j be an input to the SFEB, then

$$f_{sr} = c^{1\times1}(MSAB(GELU(c^{1\times1}(j)))) + j \tag{7.9}$$

Frequency Feature Extractor Block (FFEB). Underwater images often lose high-frequency details due to light scattering, so the FFEB can selectively emphasize or reconstruct these high-frequency components, which are crucial for super-resolution tasks. The FFEB follows a traditional architecture given by [77]. First, it employs a 2D FFT in the spatial features to convert into the frequency domain efficiently. Next, it applies a 1×1 convolution, GELU, and another 1×1 convolution to the 2D FFT features. Lastly, an inverse 2D FFT converts the processed frequency domain information back into the spatial domain, allowing for analysis of features after frequency domain operations. Let FFEB take j as input, then

$$f_{fr} = IFFT(c^{1\times1}(GELU(c^{1\times1}(FFT(j))))) \tag{7.10}$$

7.2.1.2 Prompt-guided Cross Attention Block (PCAB)

Various prompt learning strategies have been studied for both NLP and vision tasks. Prompting-based strategies are beneficial as they efficiently encode task-specific contextual details. This approach uses prompts to encode degraded information, guiding the super-resolution network. In the proposed PCAB, prompt components p_c are learnable parameters that interact with incoming multi-scale features to embed degradation information. To calculate cross-attention from multi-scale input features, PCAB first applies the 1×1 convolution followed by a 3×3 depth-wise convolution ($D_c^{3\times3}$). Next, PCAB calibrates the multi-scale input features by injecting learnable prompt components. The overall process of PCAB is defined as:

$$\begin{aligned} Q_p = D_c^{3\times3}(c^{1\times1}(f_s^1)) \odot p_c, \quad K_p = D_c^{3\times3}(c^{1\times1}(f_s^2)) \odot p_c \\ V_p = K_p, \quad f_{ca}^p = CA_c(\mathcal{R}(Q_p), \mathcal{R}(K_p)).\mathcal{R}(V_p) + f_s^1 \end{aligned} \tag{7.11}$$

where $\odot$ indicates the Hadamard product. f_s^1 and f_s^2 are the features coming from two different scales.

Table 7.4: *Quantitative comparison of the proposed approach with the state-of-the-art approaches on UFO-120 and USR-248 datasets. Red and blue indicate the best and second best results.*

Methods	Params (M)↓			FLOPs (G)↓			UFO-120						USR-248					
							PSNR↑			SSIM↑			PSNR↑			SSIM↑		
	×2	×4	×8	×2	×4	×8	×2	×3	×4	×2	×3	×4	×2	×4	×8	×2	×4	×8
SRResNet [196]	1.59	1.59	1.59	222.37	85.49	51.28	25.23	23.85	19.13	0.74	0.68	0.56	25.98	24.15	19.26	0.72	0.66	0.55
SRCNN [197]	0.06	0.06	0.06	21.30	21.30	21.30	24.75	22.22	19.05	0.72	0.65	0.56	26.81	23.38	19.97	0.76	0.67	0.57
SRGAN [196]	5.95	5.95	5.95	377.76	529.36	567.88	26.11	23.87	21.08	0.75	0.70	0.58	28.05	24.76	20.14	0.78	0.69	0.60
SRDRM [108]	0.89	1.90	2.97	203.91	291.73	313.68	24.62	-	23.15	0.72	-	0.67	28.36	24.64	21.20	0.80	0.68	0.60
SRDRM-GAN [108]	11.31	12.38	13.45	289.38	377.20	399.15	24.61	-	23.26	0.72	-	0.67	28.55	24.62	20.25	0.81	0.69	0.61
LatticeNet [201]	0.76	1.90	2.97	-	-	-	25.86	26.13	25.10	0.71	0.71	0.70	29.47	24.64	21.20	0.80	0.68	0.60
AMPCNet [110]	1.15	1.17	1.25	-	-	-	25.24	25.73	24.70	0.71	0.70	0.70	29.54	25.90	23.83	0.80	0.66	0.62
DWNet [72]	0.28	0.29	-	21.49	5.59	-	25.71	25.23	25.08	0.77	0.76	0.74	27.86	25.93	-	0.76	0.68	-
RDLN [202]	0.84	0.84	0.84	-	-	-	25.96	26.55	25.37	0.76	0.74	0.73	29.96	26.16	23.91	0.83	0.66	0.54
AFSINet [203]	2.27	2.68	4.32	169.62	50.42	20.62	26.33	25.65	25.31	0.78	0.76	0.74	29.70	26.39	23.50	0.81	0.70	0.56
MDSRDN [112]	0.29	0.32	0.34	103.4	25.95	7.95	26.04	26.41	25.53	0.78	0.75	0.73	30.57	26.38	24.16	0.83	0.71	0.62
Proposed	0.16	0.18	0.22	8.6	3.07	1.25	27.55	27.47	27.20	0.78	0.77	0.75	31.49	27.67	25.34	0.86	0.72	0.62

7.2.2 Experiments

The effectiveness of the proposed model was assessed using two widely used UISR datasets: (a) UFO-120 [109] and (b) USR-248 [108]. Data augmentation was performed during training by transforming the images with random horizontal flips and 90/180/270-degree rotations. Randomly cropped LR patches of size 64 × 64 were input during each training iteration. The PSNR and SSIM were used to evaluate the model. The model was trained using PyTorch with a batch size of 32 and a learning rate of 2e-4 on the Nvidia A100 GPU for 300 epochs.

7.2.2.1 Quantitative Results

The quantitative result of UISR on UFO-120 and USR-248 is shown in Table 7.4. The proposed model outperforms the traditional CNN-based method SRResNet by achieving notable improvements in PSNR of 2.32 dB, 3.62 dB, and 8.07 dB at ×2, ×3, and ×4 while using fewer parameters on the UFO-120 dataset. Furthermore, compared to the recent method MDSRDN, the proposed Efficient-USR gains SSIM improvements of 3.61% and 1.40% at ×2 and ×4 scale, respectively, with an 81.25% and 77.78% decrease in the number of parameters on the USR-248 dataset. All these results show the effectiveness of the method.

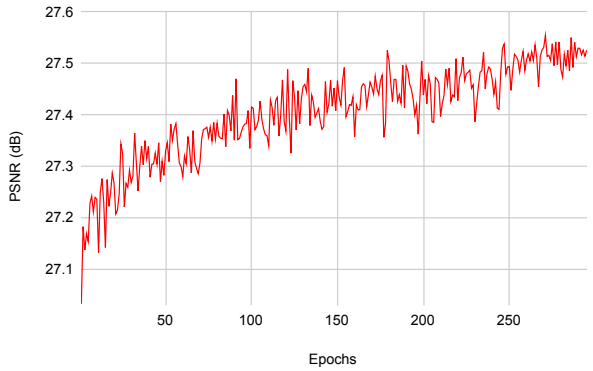


Figure 7.11: *PSNR value in each epoch on UFO-120 (×2).*

7.2.2.2 Qualitative Results

Qualitative results with other methods are provided in Figure 7.10. From Figure 7.10, it is evident that the model is capable of restoring more detailed textures and sharper edges. On the other hand, the outcomes of different models seem to be blurry or of poor quality. This qualitative comparison demonstrates that the model outperforms the high-resolution image restoration methods. Figure 7.11 shows the convergence of the model.

7.2.2.3 Ablation Study

Performed the ablation studies on the following aspects:

- **AB1-** ***Effectiveness of spatial-frequency interaction:*** The SFIB mechanism combines spatial and frequency information to enhance UISR. Experiments are conducted with and without frequency information to validate its effectiveness.
- **AB2-** ***Effectiveness of prompt-guided cross-attention*****:** Tested the effectiveness of prompt-guided cross-attention by removing prompt components from the PCA block.
- **AB3-** ***Effect of multi-scale features*****:** To validate the effectiveness of the multi-scale features, an experiment is performed on a single scale. In this case, the prompt-guided cross-attention has been changed to prompt-guided self-attention.

Table 7.5 shows the effectiveness of the above configuration in Efficient-USR on the UFO-120 (2×) dataset. Figure 7.12 exhibits that the full model addresses color and contrast challenges, improving the quality of super-resolved underwater images. The model restored more precise edges and shapes and better captured structural information.

Table 7.5: *Effect of different components of the proposed Efficient-USR.*

Model	SSIM	PSNR
AB1	0.76	26.98
AB2	0.77	27.03
AB3	0.76	26.86
Full Model	0.78	27.55

Figure 7.12: *Visual comparison of different settings of Efficient-USR.*

7.3 Summary

This contributory chapter presented two works for UISR. In the first work, a novel DL-based model is proposed for UISR by analyzing the spatial as well as frequency domain information. Various convolutional kernel sizes were utilized to capture both local and global features effectively. Additionally, an attention mechanism was incorporated to intelligently refine the features obtained, resulting in a significant improvement in the performance of the proposed model. Furthermore, it was demonstrated that using a color channel-specific L1 loss function enhances performance compared to the traditional L1 loss function. The experimental results indicate that the proposed method outperforms most of the top-published UISR schemes in recent times.

The second work proposed a lightweight model, Efficient-USR, by leveraging spatial-frequency feature representation for underwater image super-resolution. Cross-attention was utilized in both the channel and spatial dimensions for message passing between spatial and frequency features. Additionally, multi-scale features were analyzed using learnable visual prompt-guided cross-attention. The results from extensive experiments on two UISR datasets indicate that Efficient-USR has achieved state-of-the-art performance, demonstrating its superiority.

The next chapter will conclude the thesis by briefly summarizing the proposed works for enhancement and super-resolution tasks, as well as outlining future directions.

The first and second works presented in this chapter have been published at the ICASSP 2024 and ICASSP 2025 conferences, respectively.

"Simplicity is prerequisite for reliability."

~Edsger W. Dijkstra

8

Conclusion and Future Work

The primary objective of this dissertation is to propose different underwater image enhancement and super-resolution models that effectively restore degraded images without compromising visual quality. This chapter presents a summary of the eight main contributions of the thesis, organized across five chapters, and discusses potential directions for future research.

8.1 Conclusion

In the first contributory chapter, two works (Lit-Net and X-CAUNET) are discussed that utilize the color channel-aware feature learning for UIE. LitNet captured the spatial and semantic context to enhance the degraded underwater images using a multi-resolution and

multi-scale attention-based CNN model. To achieve multi-resolution analysis, different receptive fields have been applied to each color channel based on the wavelength of light. Further, an encoder-decoder module is used for multi-scale analysis. Finally, the model is optimized using a weighted color-channel-specific L1 loss. Next work X-CAUNET demonstrated the advantage of modeling the local context from individual color channels of an underwater image and combining it with global attention cues from the whole image via learnable weights.

The second contributory chapter utilized dual-color features and dual-domain features to enhance the quality of the degraded images. The first work proposed UEnhancer, which incorporated the dual-color (RGB-HSV) features to capture complementary information across different color spaces within a transformer-based hierarchical encoder-decoder model. The proposed model aims to minimize a multi-scale frequency-based SSIM loss to maintain the local structure, contrast in high-frequency regions, and retain intricate detail information. The second work introduced D2Mamba, which integrated spatial-frequency information with SSM. The SSM used a physics-based Geodesic Information-Field Heuristic scan guided by an A* search for feature traversal. Further, a Spectral Wasserstein Attenuation Loss is introduced to enforce distributional alignment in the spectral domain, enabling perceptually consistent and physically consistent color restoration in enhanced underwater images.

The third contributory chapter presented a multi-degraded representation learning-based multi-domain feature swapping model for enhancing multi-degraded underwater images. First, it generated a hazy, blurry, low-contrast, bluish, and greenish degraded image simultaneously and then performed enhancement to improve resilience for practical conditions.

In the fourth contributory chapter, the River-GEM framework is proposed to address two main limitations: (1) the lack of datasets that accurately represent the muddy water conditions found in river environments and (2) the limited effectiveness of current enhancement models in dealing with the unique challenges of muddy water images.

The final contributory chapter presents two works for UISR. The first work focused on in-

tegrating spatial features with wavelet coefficients to recover fine textures while maintaining color consistency. To provide compatibility with resource-constrained systems such as AUVs and ROVs, the second work proposed Efficient-USR. This work incorporates prompt-guided cross-attention and spatial-frequency interaction, achieving high-quality super-resolution while maintaining a lightweight model.

Extensive experiments on both UIE and UISR show that the proposed models outperform the SOTA methods. An ablation study for each work was also provided to justify the contribution of different modules. It has been experimentally demonstrated that images enhanced with the proposed models achieve significant improvements in the performance of various downstream tasks, such as underwater object detection, semantic segmentation, edge detection, and keypoint matching.

Overall, the works presented in this dissertation make significant advancements in the field of underwater vision, particularly in enhancement and super-resolution.

8.2 Future Work

The following are some additional directions in which the current dissertation study can be expanded:

- **Developing Cross-Domain Generalizable UIE and UISR Models:** Although the proposed models perform well on oceanic and riverine images, real underwater environments vary greatly across geographical regions, seasons, and turbidity conditions. Future work can focus on building cross-domain generalizable architectures that can adapt to unseen water types without retraining. This may include: (a) Domain-invariant feature learning, (b) Test-time adaptation strategies, (c) Large-scale cross-region underwater datasets, and (d) Foundation-model-style underwater vision backbones. Such models would enable robust performance across oceans, rivers, lakes, estuaries, and man-made water bodies.

- **Extending Enhancement Models to Multi-Task Underwater Vision:** En-

hancement and super-resolution models can be extended beyond image quality improvement to support downstream underwater vision tasks, such as: (a) Object detection (marine debris, fish, coral species), (b) Semantic segmentation, (c) Depth estimation, and (d) SLAM and navigation for AUVs/ROVs. Designing task-aware enhancement networks where enhancement is optimized jointly with recognition can create end-to-end pipelines for underwater robotics, monitoring, and environmental analysis.

List of Publications Related to Thesis

Journals

1. **A. Pramanick**, A. Sur, and V. V. Saradhi. "Harnessing Multi-resolution and Multi-scale Attention for Underwater Image Restoration." The Visual Computer. pp 8235–8254. 2025. [**Chapter 3**] (*refer to Section 3.1*)

2. **A. Pramanick**, A. Sur, and V. V. Saradhi. "UEnhancer: Spatially Enriched Transformer with Dual-color Information for Underwater Image Enhancement." ACM Transactions on Multimedia Computing Communications and Applications. **[Major-revision]** [**Chapter 4**] (*refer to Section 4.1*)

3. **A. Pramanick**, A. Daydar, S. Kumar, A. Sur, and V. V. Saradhi. "A Prompt-Guided Unified Model For Enhancement Of Multi-Degraded Underwater Image." IEEE Journal of Oceanic Engineering. pp. 1–16. 2026. [**Chapter 5**] (*refer to Section 5*)

Conferences

1. **A. Pramanick**, S. Sarma, and A. Sur. "X-CAUNET: Cross-Color Channel Attention with Underwater Image-Enhancing Transformer." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3550–3554. 2024. [**Chapter 3**] (*refer to Section 3.2*)

2. **A. Pramanick**, S. Roy, and A. Sur. "D2Mamba: Dual Domain Guided Informed Search in State Space Model for Underwater Image Enhancement." In Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7126–7136. 2026. [**Chapter 4**] (*refer to Section 4.2*)

3. **A. Pramanick**, C. S. Satwik, S. Kumar, A. Daydar, and A. Sur. "River-GEM: Generating and Enhancing Muddy Water Images." In IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). pp. 1–5. 2025. [**Chapter 6**] (*refer to Section 6*)

4. **A. Pramanick**, D. Megha, and A. Sur. "Attention-Based Spatial-Frequency Information Network for Underwater Single Image Super-Resolution." In IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). pp. 3560–3564. 2024. [**Chapter 7**] (*refer to Section 7.1*)

5. **A. Pramanick**, U. Bheda, and A. Sur. "Efficient-USR: Prompt Guided Dual-Domain Feature Information for Efficient Underwater Image Super-Resolution." In IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). pp. 1–5. 2025. [**Chapter 7**] (*refer to Section 7.2*)

List of Other Publications

Journals

1. **A. Pramanick**, M. Bansal, U. Srivastava, S. Ghosh, and A. Sur. "Trans-defense: Transformer-based Denoiser for Adversarial Defense with Spatial-Frequency Domain Representation." Machine Vision and Applications. **[Under-review]**

2. S. Soor, **A. Pramanick**, and A. Sur. "A Generative Adversarial Approach to Adversarial Attacks Guided by Contrastive Language-Image Pre-trained Model." SN Computer Science. **[Under-review]**

3. S. Kumar, A. S. Akshith, **A. Pramanick**, A. Sur, and R. D. Baruah. "PseudoCL: Exploring Pseudo-supervised Contrastive Learning for Unsupervised Semantic Segmentation." IEEE Transactions on Artificial Intelligence. **[Major-revision]**

Conferences

1. **A. Pramanick**, U. Bheda, and A. Sur. "ML-CrAIST: Multi-scale Low-High Frequency Information-Based Cross Attention with Image Super-Resolving Transformer." In International Conference on Pattern Recognition (ICPR). pp. 291–307. 2024.

2. A. Daydar, **A. Pramanick**, A. Sur, and S. Kanagaraj. "MedCAM-OsteoCls: Medical Context Aware Multimodal Classification of Knee Osteoarthritis." In IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). pp. 1–5. 2025.

3. A. Daydar, **A. Pramanick**, A. Sur, and S. Kanagaraj. "Diffusion Based Shape-Aware Learning With Multi-Scale Context For Segmentation Of Tibiofemoral Knee Joint Tissues: An End-To-End Approach." In IEEE International Conference on Image Processing (ICIP). pp. 1834–1839. 2025.

4. U. Srivastava, S. Roy, **A. Pramanick**, and A. Sur. "TURBIT: Generating Turbid Underwater Images With Diffusion And Differential Transformers." In IEEE International Conference on Image Processing (ICIP). pp. 2742–2747. 2025.

5. S. Ghosh, R. Sharma, S. Kumar, **A. Pramanick**, A. Sur, and P. Mitra. "MUSe: A Text-Prompted Multimodal Framework for Semi-Supervised Underwater Semantic Segmentation." In ACM International Conference on Multimodal Interaction (ICMI). 2026. **[Under-review]**

References

[1] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. "CBAM: Convolutional Block Attention Module." In Proc. of the European Conference on Computer Vision (ECCV). pp. 3–19. 2018.

[2] A. Kulkarni, S. S. Phutke, S. K. Vipparthi, and S. Murala. "Multi-Medium Image Enhancement with Attentive Deformable Transformers." *IEEE Transactions on Emerging Topics in Computational Intelligence*. pp. 1–14. 2024.

[3] S.-W. Yang, L.-H. Shen, H.-H. Shuai, and K.-T. Feng. "CMAF: Cross-Modal Augmentation via Fusion for Underwater Acoustic Image Recognition." *ACM Transactions on Multimedia Computing, Communications and Applications*. vol. 20. no. 5. pp. 1–25. 2024.

[4] B. Jiang, Y. Wang, F. Kong, and J. Wang. "Multi-Autonomous Underwater Vehicle Trajectory Planning in Ocean Current Based on Hierarchical Hunting and Evolutionary Learning." *ACM Transactions on Intelligent Systems and Technology*. vol. 16. no. 5. pp. 1–26. 2025.

[5] Y. Sun, X. Huang, Y. Cui, J. Dong, X. Zhang, and T. Yao. "An Underwater Imaging Generative Adversarial Network by Simulating the Mechanism of Light Propagation in Water." *ACM Transactions on Intelligent Systems and Technology*. vol. 16. no. 2. pp. 1–18. 2025.

[6] Y. Wang, Y. Cao, J. Zhang, F. Wu, and Z.-J. Zha. "Leveraging Deep Statistics for Underwater Image Enhancement." *ACM Transactions on Multimedia Computing, Communications, and Applications*. vol. 17. no. 3s. pp. 1–20. 2021.

[7] S.-B. Gao, M. Zhang, Q. Zhao, X.-S. Zhang, and Y.-J. Li. "Underwater Image Enhancement Using Adaptive Retinal Mechanisms." *IEEE Transactions on Image Processing*. vol. 28. no. 11. pp. 5580–5595. 2019.

[8] R. Liu, Z. Jiang, S. Yang, and X. Fan. "Twin Adversarial Contrastive Learning for Underwater Image Enhancement and Beyond." *IEEE Transactions on Image Processing*. vol. 31. pp. 4922–4936. 2022.

[9] P. Zhuang, C. Li, and J. Wu. "Bayesian Retinex Underwater Image Enhancement." *Engineering Applications of Artificial Intelligence*. vol. 101. p. 104171. 2021.

[10] M. Yang, J. Hu, C. Li, G. Rohde, Y. Du, and K. Hu. "An in-Depth Survey of Underwater Image Enhancement and Restoration." *IEEE Access*. vol. 7. pp. 123 638–123 657. 2019.

[11] B. Rokh, A. Azarpeyvand, and A. Khanteymoori. "A Comprehensive Survey on Model Quantization for Deep Neural Networks in Image Classification." *ACM Transactions on Intelligent Systems and Technology*. vol. 14. no. 6. pp. 1–50. 2023.

[12] S. Xu, M. Zhang, W. Song, H. Mei, Q. He, and A. Liotta. "A Systematic Review and Analysis of Deep Learning-Based Underwater Object Detection." *Neurocomputing*. vol. 527. pp. 204–232. 2023.

[13] D. Yi, H. B. Ahmedov, S. Jiang, Y. Li, S. J. Flinn, and P. G. Fernandes. "Coordinate-Aware Mask R-CNN with Group Normalization: A Underwater Marine Animal Instance Segmentation Framework." *Neurocomputing*. vol. 583. p. 127488. 2024.

[14] Z. Jin, M. Z. Iqbal, D. Bobkov, W. Zou, X. Li, and E. Steinbach. "A Flexible Deep CNN Framework for Image Restoration." *IEEE Transactions on Multimedia*. vol. 22. no. 4. pp. 1055–1068. 2019.

[15] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin. "Coarse-to-Fine CNN for Image Super-Resolution." *IEEE Transactions on Multimedia*. vol. 23. pp. 1489–1502. 2020.

[16] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao. "Deep Learning for Single Image Super-Resolution: A Brief Review." *IEEE Transactions on Multimedia*. vol. 21. no. 12. pp. 3106–3121. 2019.

[17] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. "Residual Dense Network for Image Super-Resolution." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2472–2481. 2018.

[18] C. Li, S. Anwar, and F. Porikli. "Underwater Scene Prior Inspired Deep Underwater Image and Video Enhancement." *Pattern Recognition*. vol. 98. p. 107038. 2020.

[19] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson. "WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images." *IEEE Robotics and Automation Letters*. vol. 3. no. 1. pp. 387–394. 2017.

[20] P. M. Uplavikar, Z. Wu, and Z. Wang. "All-in-One Underwater Image Enhancement Using Domain-Adversarial Learning." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1–8. 2019.

[21] S. Wu, T. Luo, G. Jiang, M. Yu, H. Xu, Z. Zhu, and Y. Song. "A Two-Stage Underwater Enhancement Network Based on Structure Decomposition and Characteristics of Underwater Imaging." *IEEE Journal of Oceanic Engineering*. vol. 46. no. 4. pp. 1213–1227. 2021.

[22] H. F. Tolie, J. Ren, and E. Elyan. "DICAM: Deep Inception and Channel-Wise Attention Modules for Underwater Image Enhancement." *Neurocomputing*. vol. 584. p. 127585. 2024.

[23] Y. Tao, J. Tang, X. Zhao, C. Zhou, C. Wang, and Z. Zhao. "Multi-Scale Network with Attention Mechanism for Underwater Image Enhancement." *Neurocomputing*. p. 127926. 2024.

[24] Y. Lai, H. Xu, C. Lin, T. Luo, and L. Wang. "A Two-Stage and Two-Branch Generative Adversarial Network-Based Underwater Image Enhancement." *The Visual Computer*. vol. 39. no. 9. pp. 4133–4147. 2023.

[25] J. Zhang, L. Zhu, L. Xu, and Q. Xie. "Research on the Correlation Between Image Enhancement and Underwater Object Detection." In Proc. of the Chinese Automation Congress (CAC). pp. 5928–5933. 2020.

[26] J. M. Goggin. "Underwater Archaeology: Its Nature and Limitations." *American Antiquity*. vol. 25. no. 3. p. 348–354. 1960.

[27] N. A. Ubina and S.-C. Cheng. "A Review of Unmanned System Technologies with Its Application to Aquaculture Farm Monitoring and Management." *Drones*. vol. 6. no. 1. 2022.

[28] R. Stopforth, S. Holtzhausen, G. Bright, N. S. Tlale, and C. M. Kumile. "Robots for Search and Rescue Purposes in Urban and Underwater Environments - A Survey and Comparison." In Proc. of the International Conference on Mechatronics and Machine Vision in Practice (ICMMVP). pp. 476–480. 2008.

[29] N. Qiao and L. Di. "Underwater Image Enhancement Combining Low-Dimensional and Global Features." *The Visual Computer*. vol. 39. no. 7. pp. 3029–3039. 2023.

[30] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao. "An Underwater Image Enhancement Benchmark Dataset and Beyond." *IEEE Transactions on Image Processing*. vol. 29. pp. 4376–4389. 2019.

[31] C. Fabbri, M. J. Islam, and J. Sattar. "Enhancing Underwater Imagery Using Generative Adversarial Networks." In Proc. of the IEEE International Conference on Robotics and Automation (ICRA). pp. 7159–7165. 2018.

[32] M. J. Islam, Y. Xia, and J. Sattar. "Fast Underwater Image Enhancement for Improved Visual Perception." *IEEE Robotics and Automation Letters*. vol. 5. no. 2. pp. 3227–3234. 2020.

[33] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren. "Underwater Image Enhancement via Medium Transmission-Guided Multi-Color Space Embedding." *IEEE Transactions on Image Processing*. vol. 30. pp. 4985–5000. 2021.

[34] C. Liu, X. Shu, L. Pan, J. Shi, and B. Han. "Multi-Scale Underwater Image Enhancement in RGB and HSV Color Spaces." *IEEE Transactions on Instrumentation and Measurement*. vol. 72. pp. 1–14. 2023.

[35] Q. Xue, H. Hu, Y. Bai, R. Cheng, P. Wang, and N. Song. "Underwater Image Enhancement Algorithm Based on Color Correction and Contrast Enhancement." *The Visual Computer*. vol. 40. no. 8. pp. 5475–5502. 2024.

[36] L. Saad Saoud, M. Elmezain, A. Sultan, M. Heshmat, L. Seneviratne, and I. Hussain. "Seeing Through the Haze: A Comprehensive Review of Underwater Image Enhancement Techniques." *IEEE Access*. vol. 12. pp. 145 206–145 233. 2024.

[37] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li. "Underwater Image Enhancement via Minimal Color Loss and Locally Adaptive Contrast Enhancement." *IEEE Transactions on Image Processing*. vol. 31. pp. 3997–4010. 2022.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems*. vol. 25. 2012.

[39] F. Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1251–1258. 2017.

[40] Y. Li, X. Zhang, and D. Chen. "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1091–1100. 2018.

[41] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng. "Strip Pooling: Rethinking Spatial Pooling for Scene Parsing." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4003–4012. 2020.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention Is All You Need." *Advances in Neural Information Processing Systems*. vol. 30. 2017.

[43] A. Gu, K. Goel, and C. Ré. "Efficiently Modeling Long Sequences with Structured State Spaces." *arXiv preprint arXiv:2111.00396*. 2021.

[44] D. Akkaynak and T. Treibitz. "Sea-Thru: A Method for Removing Water From Underwater Images." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1682–1691. 2019.

[45] D. Berman, D. Levy, S. Avidan, and T. Treibitz. "Underwater Single Image Color Restoration Using Haze-Lines and a New Quantitative Dataset." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 43. no. 8. pp. 2822–2837. 2020.

[46] J. Y. Chiang and Y.-C. Chen. "Underwater Image Enhancement by Wavelength Compensation and Dehazing." *IEEE Transactions on Image Processing*. vol. 21. no. 4. pp. 1756–1769. 2011.

[47] P. L. Drews, E. R. Nascimento, S. S. Botelho, and M. F. M. Campos. "Underwater Depth Estimation and Image Restoration Based on Single Images." *IEEE Computer Graphics and Applications*. vol. 36. no. 2. pp. 24–35. 2016.

[48] C.-Y. Li, J.-C. Guo, R.-M. Cong, Y.-W. Pang, and B. Wang. "Underwater Image Enhancement by Dehazing with Minimum Information Loss and Histogram Distribution Prior." *IEEE Transactions on Image Processing*. vol. 25. no. 12. pp. 5664–5677. 2016.

[49] C. Li, J. Quo, Y. Pang, S. Chen, and J. Wang. "Single Underwater Image Restoration by Blue-Green Channels Dehazing and Red Channel Correction." In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1731–1735. 2016.

[50] Y.-T. Peng and P. C. Cosman. "Underwater Image Restoration Based on Image Blurriness and Light Absorption." *IEEE Transactions on Image Processing*. vol. 26. no. 4. pp. 1579–1594. 2017.

[51] W. Song, Y. Wang, D. Huang, A. Liotta, and C. Perra. "Enhancement of Underwater Images with Statistical Model of Background Light and Optimization of Transmission Map." *IEEE Transactions on Broadcasting*. vol. 66. no. 1. pp. 153–169. 2020.

[52] J. Xie, G. Hou, G. Wang, and Z. Pan. "A Variational Framework for Underwater Image Dehazing and Deblurring." *IEEE Transactions on Circuits and Systems for Video Technology*. vol. 32. no. 6. pp. 3514–3526. 2021.

[53] G. Hou, J. Li, G. Wang, H. Yang, B. Huang, and Z. Pan. "A Novel Dark Channel Prior Guided Variational Framework for Underwater Image Restoration." *Journal of Visual Communication and Image Representation*. vol. 66. p. 102732. 2020.

[54] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila. "Automatic Red-Channel Underwater Image Restoration." *Journal of Visual Communication and Image Representation*. vol. 26. pp. 132–145. 2015.

[55] D. Berman, T. Treibitz, and S. Avidan. "Diving Into Haze-Lines: Color Restoration of Underwater Images." In Proc. of the British Machine Vision Conference (BMVC). 2017.

[56] W. Song, Y. Wang, D. Huang, and D. Tjondronegoro. "A Rapid Scene Depth Estimation Model Based on Underwater Light Attenuation Prior for Underwater Image Restoration." In Proc. of the Pacific-Rim Conference on Multimedia (PCM). pp. 678–688. 2018.

[57] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert. "Color Balance and Fusion for Underwater Image Enhancement." *IEEE Transactions on Image Processing*. vol. 27. no. 1. pp. 379–393. 2017.

[58] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and M. Sbert. "Color Channel Compensation (3C): A Fundamental Pre-Processing Step for Image Enhancement." *IEEE Transactions on Image Processing*. vol. 29. pp. 2653–2665. 2019.

[59] D. Huang, Y. Wang, W. Song, J. Sequeira, and S. Mavromatis. "Shallow-Water Image Enhancement Using Relative Global Histogram Stretching Based on Adaptive Parameter Acquisition." In Proc. of the International Conference on MultiMedia Modeling (MMM). pp. 453–465. 2018.

[60] B.-H. Chen, S.-C. Huang, and J. H. Ye. "Hazy Image Restoration by Bi-Histogram Modification." *ACM Transactions on Intelligent Systems and Technology*. vol. 6. no. 4. 2015.

[61] R. Hummel. "Image enhancement by histogram transformation." *Computer Graphics and Image Processing*. vol. 6. no. 2. pp. 184–195. 1977.

[62] X. Li, G. Hou, K. Li, and Z. Pan. "Enhancing Underwater Image via Adaptive Color and Contrast Enhancement, and Denoising." *Engineering Applications of Artificial Intelligence*. vol. 111. p. 104759. 2022.

[63] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding. "A Retinex-Based Enhancing Approach for Single Underwater Image." In Proc. of the IEEE International Conference on Image Processing (ICIP). pp. 4572–4576. 2014.

[64] K. Jiang, Q. Wang, Z. An, Z. Wang, C. Zhang, and C.-W. Lin. "Mutual Retinex: Combining Transformer and CNN for Image Enhancement." *IEEE Transactions on Emerging Topics in Computational Intelligence*. vol. 8. no. 3. pp. 2240–2252. 2024.

[65] S. Zhang, T. Wang, J. Dong, and H. Yu. "Underwater Image Enhancement via Extended Multi-Scale Retinex." *Neurocomputing*. vol. 245. pp. 1–9. 2017.

[66] P. Zhuang. "Retinex Underwater Image Enhancement with Multiorder Gradient Priors." In Proc. of the IEEE International Conference on Image Processing (ICIP). pp. 1709–1713. 2021.

[67] P. Zhuang and X. Ding. "Underwater Image Enhancement Using an Edge-Preserving Filtering Retinex Algorithm." *Multimedia Tools and Applications*. vol. 79. pp. 17 257–17 277. 2020.

[68] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert. "Enhancing Underwater Images and Videos by Fusion." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 81–88. 2012.

[69] X. Li, G. Hou, L. Tan, and W. Liu. "A Hybrid Framework for Underwater Image Enhancement." *IEEE Access*. vol. 8. pp. 197 448–197 462. 2020.

[70] N. Shashar, R. T. Hanlon, and A. d. Petz. "Polarization Vision Helps Detect Transparent Prey." *Nature*. vol. 393. no. 6682. pp. 222–223. 1998.

[71] T. Treibitz and Y. Y. Schechner. "Turbid Scene Enhancement Using Multi-Directional Illumination Fusion." *IEEE Transactions on Image Processing*. vol. 21. no. 11. pp. 4662–4667. 2012.

[72] P. Sharma, I. Bisht, and A. Sur. "Wavelength-Based Attributed Deep Neural Network for Underwater Image Restoration." *ACM Transactions on Multimedia Computing, Communications and Applications*. vol. 19. no. 1. pp. 1–23. 2023.

[73] A. Pramanick, S. Sarma, and A. Sur. "X-Caunet: Cross-Color Channel Attention with Underwater Image-Enhancing Transformer." In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3550–3554. 2024.

[74] L. Peng, C. Zhu, and L. Bian. "U-Shape Transformer for Underwater Image Enhancement." *IEEE Transactions on Image Processing*. vol. 32. pp. 3066–3079. 2023.

[75] Q. Qi, K. Li, H. Zheng, X. Gao, G. Hou, and K. Sun. "SGUIE-Net: Semantic Attention Guided Underwater Image Enhancement with Multi-Scale Perception." *IEEE Transactions on Image Processing*. vol. 31. pp. 6816–6830. 2022.

[76] P. Mishra, S. K. Vipparthi, and S. Murala. "U-Enhance: Underwater Image Enhancement Using Wavelet Triple Self-Attention." In Proc. of the Asian Conference on Computer Vision (ACCV). pp. 84–101. 2024.

[77] R. Khan, P. Mishra, N. Mehta, S. S. Phutke, S. K. Vipparthi, S. Nandi, and S. Murala. "Spectroformer: Multi-Domain Query Cascaded Transformer Network for Underwater Image Enhancement." In Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1454–1463. 2024.

[78] A. Chandrasekar, M. Sreenivas, and S. Biswas. "PhISH-Net: Physics-Inspired System for High-Resolution Underwater Image Enhancement." In Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1506–1516. 2024.

[79] C. Zhao, W. Cai, C. Dong, and C. Hu. "Wavelet-Based Fourier Information Interaction with Frequency Diffusion Adjustment for Underwater Image Restoration." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8281–8291. 2024.

[80] Y. Zhang, Q. Jiang, P. Liu, S. Gao, X. Pan, and C. Zhang. "Underwater Image Enhancement Using Deep Transfer Learning Based on a Color Restoration Model." *IEEE Journal of Oceanic Engineering*. vol. 48. no. 2. pp. 489–514. 2023.

[81] A. Naik, A. Swarnakar, and K. Mittal. "Shallow-UWNet: Compressed Model for Underwater Image Enhancement." In Proc. of the AAAI Conference on Artificial Intelligence. vol. 35. no. 18. pp. 15 853–15 854. 2021.

[82] P. Liu, G. Wang, H. Qi, C. Zhang, H. Zheng, and Z. Yu. "Underwater Image Enhancement with a Deep Residual Framework." *IEEE Access*. vol. 7. pp. 94 614–94 629. 2019.

[83] J. Wu, X. Liu, N. Qin, Q. Lu, and X. Zhu. "Two-Stage Progressive Underwater Image Enhancement." *IEEE Transactions on Instrumentation and Measurement*. vol. 73. pp. 1–18. 2024.

[84] H.-H. Yang, K.-C. Huang, and W.-T. Chen. "LAFFNet: A Lightweight Adaptive Feature Fusion Network for Underwater Image Enhancement." In Proc. of the IEEE International Conference on Robotics and Automation (ICRA). pp. 685–692. 2021.

[85] C. Li, J. Guo, and C. Guo. "Emerging From Water: Underwater Image Color Correction Based on Weakly Supervised Color Transfer." *IEEE Signal Processing Letters*. vol. 25. no. 3. pp. 323–327. 2018.

[86] Y. Guo, H. Li, and P. Zhuang. "Underwater Image Enhancement Using a Multiscale Dense Generative Adversarial Network." *IEEE Journal of Oceanic Engineering*. vol. 45. no. 3. pp. 862–870. 2019.

[87] C. Desai, R. A. Tabib, S. S. Reddy, U. Patil, and U. Mudenagudi. "RUIG: Realistic Underwater Image Generation Towards Restoration." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2181–2189. 2021.

[88] X. Yang, H. Li, and R. Chen. "Underwater Image Enhancement with Image Colorfulness Measure." *Signal Processing: Image Communication*. vol. 95. p. 116225. 2021.

[89] P. Hambarde, S. Murala, and A. Dhall. "UW-GAN: Single-Image Depth Estimation and Image Enhancement for Underwater Images." *IEEE Transactions on Instrumentation and Measurement*. vol. 70. pp. 1–12. 2021.

[90] Z. Huang, J. Li, Z. Hua, and L. Fan. "Underwater Image Enhancement via Adaptive Group Attention-Based Multiscale Cascade Transformer." *IEEE Transactions on Instrumentation and Measurement*. vol. 71. pp. 1–18. 2022.

[91] Y. Zhou, H. Xu, G. Jiang, M. Yu, Y. Chen, and T. Luo. "UIE-SFIFormer: Underwater Image Enhancement Based on Physical-Guided Spatial–Frequency Interaction Transformer." *IEEE Journal of Oceanic Engineering*. vol. 50. no. 2. pp. 727–742. 2024.

[92] J. Qu, X. Cao, S. Jiang, J. You, and Z. Yu. "UIEFormer: Lightweight Vision Transformer for Underwater Image Enhancement." *IEEE Journal of Oceanic Engineering*. vol. 50. no. 2. pp. 851–865. 2025.

[93] Y. Qing, L. Shen, Z. Fang, and Y. Wang. "HG2Former: HSV-Gamma Guided Transformers for Efficient Underwater Image Enhancement." *IEEE Journal of Oceanic Engineering*. vol. 50. no. 2. pp. 866–878. 2025.

[94] J. D. Hamilton. "State-Space Models." *Handbook of Econometrics*. vol. 4. pp. 3039–3080. 1994.

[95] A. Gu, K. Goel, and C. Ré. "Efficiently Modeling Long Sequences with Structured State Spaces." In Proc. of the International Conference on Learning Representations (ICLR). pp. 1–27. 2022.

[96] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré. "Combining Recurrent, Convolutional, and Continuous-Time Models with Linear State Space Layers." *Advances in Neural Information Processing Systems*. vol. 34. pp. 572–585. 2021.

[97] A. Gu and T. Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces." *arXiv Preprint arXiv:2312.00752*. 2023.

[98] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi. "RSMamba: Remote Sensing Image Classification with State Space Model." *IEEE Geoscience and Remote Sensing Letters*. vol. 21. pp. 1–5. 2024.

[99] T. D. Q. Dang, H. H. Nguyen, and A. Tiulpin. "LoG-VMamba: Local-Global Vision Mamba for Medical Image Segmentation." In Proc. of the Asian Conference on Computer Vision (ACCV). pp. 548–565. 2024.

[100] G. An, A. He, Y. Wang, and J. Guo. "UWMamba: Underwater Image Enhancement with State Space Model." *IEEE Signal Processing Letters*. vol. 31. pp. 2725–2729. 2024.

[101] W.-T. Lin, Y.-X. Lin, J.-W. Chen, and K.-L. Hua. "PixMamba: Leveraging State Space Models in a Dual-Level Architecture for Underwater Image Enhancement." In Proc. of the Asian Conference on Computer Vision (ACCV). pp. 3622–3637. 2024.

[102] M. Guan, H. Xu, G. Jiang, M. Yu, Y. Chen, T. Luo, and Y. Song. "WaterMamba: Visual State Space Model for Underwater Image Enhancement." *arXiv Preprint arXiv:2405.08419*. 2024.

[103] C. Dong, C. Zhao, W. Cai, and B. Yang. "O-Mamba: O-Shape State-Space Model for Underwater Image Enhancement." *arXiv Preprint arXiv:2408.12816*. 2024.

[104] L. Peng and L. Bian. "Adaptive Dual-Domain Learning for Underwater Image Enhancement." In Proc. of the AAAI Conference on Artificial Intelligence. vol. 39. no. 6. pp. 6461–6469. 2025.

[105] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. A. Morel. "Neighbor Embedding Based Single-Image Super-Resolution Using Semi-Nonnegative Matrix Factorization." In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1289–1292. 2012.

[106] Y. Chen, K. Niu, Z. Zeng, and Y. Pan. "A Wavelet-Based Deep Learning Method for Underwater Image Super-Resolution Reconstruction." *IEEE Access*. vol. 8. pp. 117 759–117 769. 2020.

[107] H. Lu, Y. Li, S. Nakashima, H. Kim, and S. Serikawa. "Underwater Image Super-Resolution by Descattering and Fusion." *IEEE Access*. vol. 5. pp. 670–679. 2017.

[108] 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). pp. 900–906. 2020.

[109] M. J. Islam, P. Luo, and J. Sattar. "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception." *arXiv Preprint arXiv:2002.01155*. 2020.

[110] Y. Zhang, S. Yang, Y. Sun, S. Liu, and X. Li. "Attention-Guided Multi-Path Cross-CNN for Underwater Image Super-Resolution." *Signal, Image and Video Processing*. vol. 16. no. 1. pp. 155–163. 2022.

[111] T. Ren, H. Xu, G. Jiang, M. Yu, X. Zhang, B. Wang, and T. Luo. "Reinforced Swin-Convs Transformer for Simultaneous Underwater Sensing Scene Image Enhancement and Super-Resolution." *IEEE Transactions on Geoscience and Remote Sensing*. vol. 60. pp. 1–16. 2022.

[112] B. Liu, X. Ning, S. Ma, and Y. Yang. "Multi-Scale Dense Spatially-Adaptive Residual Distillation Network for Lightweight Underwater Image Super-Resolution." *Frontiers in Marine Science*. vol. 10. p. 1328436. 2024.

[113] A. Hore and D. Ziou. "Image Quality Metrics: PSNR vs. SSIM." In Proc. of the International Conference on Pattern Recognition (ICPR). pp. 2366–2369. 2010.

[114] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*. vol. 13. no. 4. pp. 600–612. 2004.

[115] Z. Wang, E. P. Simoncelli, and A. C. Bovik. "Multiscale Structural Similarity for Image Quality Assessment." In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2. pp. 1398–1402. 2003.

[116] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595. 2018.

[117] K. Panetta, C. Gao, and S. Agaian. "Human-Visual-System-Inspired Underwater Image Quality Measures." *IEEE Journal of Oceanic Engineering*. vol. 41. no. 3. pp. 541–551. 2015.

[118] A. Mittal, A. K. Moorthy, and A. C. Bovik. "No-Reference Image Quality Assessment in the Spatial Domain." *IEEE Transactions on Image Processing*. vol. 21. no. 12. pp. 4695–4708. 2012.

[119] Z. Chen, G. Qiu, P. Li, L. Zhu, X. Yang, and B. Sheng. "MNGNAS: Distilling Adaptive Combination of Multiple Searched Networks for One-Shot Neural Architecture Search." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 45. no. 11. pp. 13 489–13 508. 2023.

[120] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going Deeper with Convolutions." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9. 2015.

[121] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In Proc. of the International Conference on Machine Learning (ICML). pp. 448–456. 2015.

[122] J. Johnson, A. Alahi, and L. Fei-Fei. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution." In Proc. of the European Conference on Computer Vision (ECCV). pp. 694–711. 2016.

[123] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv Preprint arXiv:1409.1556*. 2014.

[124] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo. "Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light." *IEEE Transactions on Circuits and Systems for Video Technology*. vol. 30. no. 12. pp. 4861–4875. 2020.

[125] H. Li, J. Li, and W. Wang. "A Fusion Adversarial Underwater Image Enhancement Network with a Public Test Dataset." *arXiv Preprint arXiv:1906.06819*. 2019.

[126] L. Chang, H. Song, M. Li, and M. Xiang. "UIDEF: A Real-World Underwater Image Dataset and a Color-Contrast Complementary Image Enhancement Framework." *ISPRS Journal of Photogrammetry and Remote Sensing*. vol. 196. pp. 415–428. 2023.

[127] P. Drews, E. Nascimento, F. Moraes, S. Botelho, and M. Campos. "Transmission Estimation in Underwater Single Images." In Proc. of the IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 825–830. 2013.

[128] Y.-T. Peng, K. Cao, and P. C. Cosman. "Generalization of the Dark Channel Prior for Single Image Restoration." *IEEE Transactions on Image Processing*. vol. 27. no. 6. pp. 2856–2868. 2018.

[129] Y. Wang, J. Guo, H. Gao, and H. Yue. "Uiec^2-Net: CNN-Based Underwater Image Enhancement Using Two Color Spaces." *Signal Processing: Image Communication*. vol. 96. p. 116250. 2021.

[130] Z. Fu, W. Wang, Y. Huang, X. Ding, and K.-K. Ma. "Uncertainty Inspired Underwater Image Enhancement." In Proc. of the European Conference on Computer Vision (ECCV). pp. 465–482. 2022.

[131] J. Zhou, B. Li, D. Zhang, J. Yuan, W. Zhang, Z. Cai, and J. Shi. "UGIF-Net: An Efficient Fully Guided Information Flow Network for Underwater Image Enhancement." *IEEE Transactions on Geoscience and Remote Sensing*. vol. 61. pp. 1–17. 2023.

[132] Y. Tang, H. Kawasaki, and T. Iwaguchi. "Underwater Image Enhancement by Transformer-Based Diffusion Model with Non-Uniform Sampling for Skip Strategy." In Proc. of the ACM International Conference on Multimedia. pp. 5419–5427. 2023.

[133] H. Wang, S. Sun, L. Chang, H. Li, W. Zhang, A. C. Frery, and P. Ren. "INSPIRATION: A Reinforcement Learning-Based Human Visual Perception-Driven Image Enhancement Paradigm for Underwater Scenes." *Engineering Applications of Artificial Intelligence*. vol. 133. p. 108411. 2024.

[134] Z. Ma and C. Oh. "A Wavelet-Based Dual-Stream Network for Underwater Image Enhancement." In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2769–2773. 2022.

[135] S. Anwar, C. Li, and F. Porikli. "Deep Underwater Image Enhancement." *arXiv preprint arXiv:1807.03528*. 2018.

[136] A. K. Mishra, M. Kumar, and M. S. Choudhry. "Fusion of Multiscale Gradient Domain Enhancement and Gamma Correction for Underwater Image/Video Enhancement and Restoration." *Optics and Lasers in Engineering*. vol. 178. p. 108154. 2024.

[137] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar. "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark." In Proc. of the IEEE International Conference on Intelligent Robots and Systems (IROS). pp. 1769–1776. 2020.

[138] G. Jocher, A. Chaurasia, A. Stoken *et al.* "YOLOv5." https://github.com/ultralytics/yolov5. 2020.

[139] Roboflow. "Aquarium Dataset." https://public.roboflow.com/object-detection/aquarium. 2020.

[140] A. Radford *et al.* "Learning Transferable Visual Models From Natural Language Supervision." In Proc. of the International Conference on Machine Learning (ICML). pp. 8748–8763. 2021.

[141] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. "Restormer: Efficient Transformer for High-Resolution Image Restoration." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5728–5739. 2022.

[142] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao. "Learning Enriched Features for Fast Image Restoration and Enhancement." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 45. no. 2. pp. 1934–1948. 2022.

[143] S. Anwar and C. Li. "Diving Deeper Into Underwater Image Enhancement: A Survey." *Signal Processing: Image Communication*. vol. 89. p. 115978. 2020.

[144] W.-Y. Hsu and J.-Y. Yang. "Cross-Hierarchical Structure-Detail-Aware Transformer for Single Image Deblurring." *ACM Transactions on Intelligent Systems and Technology*. 2025.

[145] S.-C. Huang, D.-W. Jaw, B.-H. Chen, and S.-Y. Kuo. "Single Image Snow Removal Using Sparse Representation and Particle Swarm Optimizer." *ACM Transactions on Intelligent Systems and Technology*. vol. 11. no. 2. 2020.

[146] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li. "Uformer: A General U-Shaped Transformer for Image Restoration." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17 683–17 693. 2022.

[147] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo. "Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions." *arXiv preprint arXiv:1901.05320*. 2019.

[148] Y. Qing, Y. Wang, H. Yan, X. Xie, and Z. Wu. "Unformer: A Transformer-Based Approach for Adaptive Multiscale Feature Aggregation in Underwater Image Enhancement." *IEEE Transactions on Artificial Intelligence*. vol. 6. no. 4. pp. 1024–1037. 2024.

[149] H. Zhang, H. Xu, X. Yu, X. Zhang, X. Gao, and C. Wu. "CDF-UIE: Leveraging Cross-Domain Fusion for Underwater Image Enhancement." *IEEE Transactions on Geoscience and Remote Sensing*. vol. 63. pp. 1–15. 2025.

[150] X. Guo, X. Chen, S. Wang, and C.-M. Pun. "Underwater image restoration through a prior guided hybrid sense approach and extensive benchmark analysis." *IEEE Transactions on Circuits and Systems for Video Technology*. vol. 35. no. 5. pp. 4784–4800. 2025.

[151] B. Wang, H. Xu, G. Jiang, M. Yu, T. Ren, T. Luo, and Z. Zhu. "UIE-Convformer: Underwater Image Enhancement based on Convolution and Feature Fusion Transformer." *IEEE Transactions on Emerging Topics in Computational Intelligence*. vol. 8. no. 2. pp. 1952–1968. 2024.

[152] Z. Wang, L. Shen, M. Xu, M. Yu, K. Wang, and Y. Lin. "Domain Adaptation for Underwater Image Enhancement." *IEEE Transactions on Image Processing*. vol. 32. pp. 1442–1457. 2023.

[153] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar. "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark." In Proc. of the IEEE International Conference on Intelligent Robots and Systems (IROS). pp. 1769–1776. 2020.

[154] C.-M. Fan, T.-J. Liu, and K.-H. Liu. "Half Wavelet Attention on M-Net+ for Low-Light Image Enhancement." In Proc. of the IEEE International Conference on Image Processing (ICIP). pp. 3878–3882. 2022.

[155] S. Hao, X. Han, Y. Guo, and M. Wang. "Decoupled Low-Light Image Enhancement." *ACM Transactions on Multimedia Computing, Communications, and Applications*. vol. 18. no. 4. pp. 1–19. 2022.

[156] Y. Yin, D. Xu, C. Tan, P. Liu, Y. Zhao, and Y. Wei. "CLE Diffusion: Controllable Light Enhancement Diffusion Model." In Proc. of the ACM International Conference on Multimedia. pp. 8145–8156. 2023.

[157] S. Yang, D. Zhou, J. Cao, and Y. Guo. "LightingNet: An Integrated Learning Method for Low-Light Image Enhancement." *IEEE Transactions on Computational Imaging*. vol. 9. pp. 29–42. 2023.

[158] P. Shi, X. Xu, X. Fan, X. Yang, and Y. Xin. "LL-UNet++:UNet++ Based Nested Skip Connections Network for Low-Light Image Enhancement." *IEEE Transactions on Computational Imaging*. vol. 10. pp. 510–521. 2024.

[159] G. Han, K. Wu, F. Zeng, J. Liu, and S. Kwong. "Dual-Stream Adaptive Convergent Low-Light Image Enhancement Network Based on Frequency Perception." *IEEE Transactions on Computational Imaging*. vol. 9. pp. 1152–1164. 2023.

[160] L. Zhao, B. Chen, J. Zhang, A. Wang, and H. Bai. "RIRO: From Retinex-Inspired Reconstruction Optimization Model to Deep Low-Light Image Enhancement Unfolding Network." *IEEE Transactions on Computational Imaging*. vol. 10. pp. 969–983. 2024.

[161] H. Hussain, P. S. Tamizharasan, and P. K. Yadav. "OptiRet-Net: An Optimized Low-Light Image Enhancement Technique for CV-Based Applications in Resource-Constrained Environments." *ACM Transactions on Intelligent Systems and Technology*. vol. 15. no. 6. 2024.

[162] K. He, J. Sun, and X. Tang. "Single Image Haze Removal Using Dark Channel Prior." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 33. no. 12. pp. 2341–2353. 2010.

[163] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. "VMamba: Visual State Space Model." *Advances in Neural Information Processing Systems*. vol. 37. pp. 103 031–103 063. 2024.

[164] T. Dao and A. Gu. "Transformers Are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality." In Proc. of the International Conference on Machine Learning (ICML). pp. 10 041–10 071. 2024.

[165] Q. Zhu, Y. Fang, Y. Cai, C. Chen, and L. Fan. "Rethinking Scanning Strategies with Vision Mamba in Semantic Segmentation of Remote Sensing Imagery: An Experimental Study." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. vol. 17. pp. 18 223–18 234. 2024.

[166] P. E. Hart, N. J. Nilsson, and B. Raphael. "A Formal Basis for the Heuristic Determination of Minimum Cost Paths." *IEEE Transactions on Systems Science and Cybernetics*. vol. 4. no. 2. pp. 100–107. 1968.

[167] V. Caselles, R. Kimmel, and G. Sapiro. "Geodesic Active Contours." In Proc. of IEEE International Conference on Computer Vision (ICCV). pp. 694–699. 1995.

[168] A. Furnari, G. M. Farinella, A. R. Bruna, and S. Battiato. "Generalized Sobel Filters for Gradient Estimation of Distorted Images." In Proc. of the IEEE International Conference on Image Processing (ICIP). pp. 3250–3254. 2015.

[169] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang. "Multi-Scale Boosted Dehazing Network with Dense Feature Fusion." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2157–2167. 2020.

[170] Y. Wang, X. Ding, R. Wang, J. Zhang, and X. Fu. "Fusion-Based Underwater Image Enhancement by Wavelet Decomposition." In 2017 IEEE International Conference on Industrial Technology (ICIT). pp. 1013–1018. 2017.

[171] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. "Loss Functions for Image Restoration with Neural Networks." *IEEE Transactions on Computational Imaging*. vol. 3. no. 1. pp. 47–57. 2016.

[172] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." In 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. 2009.

[173] L. Rüschendorf. "The Wasserstein Distance and Approximation Theorems." *Probability Theory and Related Fields*. vol. 70. no. 1. pp. 117–129. 1985.

[174] D. Berman, T. Treibitz, and S. Avidan. "Single Image Dehazing Using Haze-Lines." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 42. no. 3. pp. 720–734. 2018.

[175] R. Pucci and N. Martinel. "CE-VAE: Capsule-Enhanced Variational Autoencoder for Underwater Image Enhancement." In Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2113–2123. 2025.

[176] M. R. Khan, A. Negi, A. Kulkarni, S. S. Phutke, S. K. Vipparthi, and S. Murala. "PhaseFormer: Phase-Based Attention Mechanism for Underwater Image Restoration

and Beyond." In Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 9618–9629. 2025.

[177] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li. "Contrastive Semi-Supervised Learning for Underwater Image Restoration via Reliable Bank." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18 145–18 155. 2023.

[178] Z. Cheng, G. Fan, J. Zhou, M. Gan, and C. L. P. Chen. "FDCE-Net: Underwater Image Enhancement with Embedding Frequency and Dual Color Encoder." *IEEE Transactions on Circuits and Systems for Video Technology*. vol. 35. no. 2. pp. 1728–1744. 2025.

[179] S. Zhou, J. Pan, J. Shi, D. Chen, L. Qu, and J. Yang. "Seeing the Unseen: A Frequency Prompt-Guided Transformer for Image Restoration." In Proc. of the European Conference on Computer Vision (ECCV). pp. 246–264. 2024.

[180] P. Vaishnav, Z. Syed Waqas, K. Salman, and K. Fahad Shahbaz. "PromptIR: Prompting for All-in-One Blind Image Restoration." *Advances in Neural Information Processing Systems*. vol. 36. pp. 71 275–71 293. 2023.

[181] Z. Wu, W. Liu, J. Wang, J. Li, and D. Huang. "FrePrompter: Frequency Self-Prompt for All-in-One Image Restoration." *Pattern Recognition*. vol. 161. p. 111223. 2025.

[182] B. Jähne. *Digital Image Processing*. Springer Science and Business Media, 2005.

[183] M. Jha and A. K. Bhandari. "CBLA: Color-Balanced Locally Adjustable Underwater Image Enhancement." *IEEE Transactions on Instrumentation and Measurement*. vol. 73. pp. 1–11. 2024.

[184] H. Feizi, M. Sattari, M. Mosaferi, and H. Apaydin. "An Image-Based Deep Learning Model for Water Turbidity Estimation in Laboratory Conditions." *International Journal of Environmental Science and Technology*. vol. 20. no. 1. pp. 149–160. 2023.

[185] Q. Jiang, Y. Kang, Z. Wang, W. Ren, and C. Li. "Perception-Driven Deep Underwater Image Enhancement Without Paired Supervision." *IEEE Transactions on Multimedia*. vol. 26. pp. 4884–4897. 2023.

[186] F. Kherif and A. Latypova. "Principal Component Analysis." pp. 209–225. 2020.

[187] G. W. Stewart. "On the Early History of the Singular Value Decomposition." *SIAM review*. vol. 35. no. 4. pp. 551–566. 1993.

[188] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10 371–10 381. 2024.

[189] J. T. Barron. "A General and Adaptive Robust Loss Function." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4331–4339. 2019.

[190] Z. Wang, W. Liu, Y. Wang, and B. Liu. "AGCycleGAN: Attention-Guided CycleGAN for Single Underwater Image Restoration." In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2779–2783. 2022.

[191] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks." In Proc. of the European Conference on Computer Vision Workshops (ECCVW). 2018.

[192] P. K. Sharma, P. Jain, and A. Sur. "Dual-Domain Single Image De-Raining Using Conditional Generative Adversarial Network." In Proc. of the IEEE International Conference on Image Processing (ICIP). pp. 2796–2800. 2019.

[193] P. Sharma, P. Jain, and A. Sur. "Scale-Aware Conditional Generative Adversarial Network for Image Dehazing." In Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2355–2365. 2020.

[194] D. Zhang. "Wavelet Transform." *Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval*. pp. 35–44. 2019.

[195] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1874–1883. 2016.

[196] C. Ledig *et al.* "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4681–4690. 2017.

[197] C. Dong, C. C. Loy, K. He, and X. Tang. "Image Super-Resolution Using Deep Convolutional Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 38. no. 2. pp. 295–307. 2015.

[198] H. Gao, J. Yang, Y. Zhang, N. Wang, J. Yang, and D. Dang. "Prompt-Based Ingredient-Oriented All-in-One Image Restoration." *IEEE Transactions on Circuits and Systems for Video Technology*. pp. 1–1. 2024.

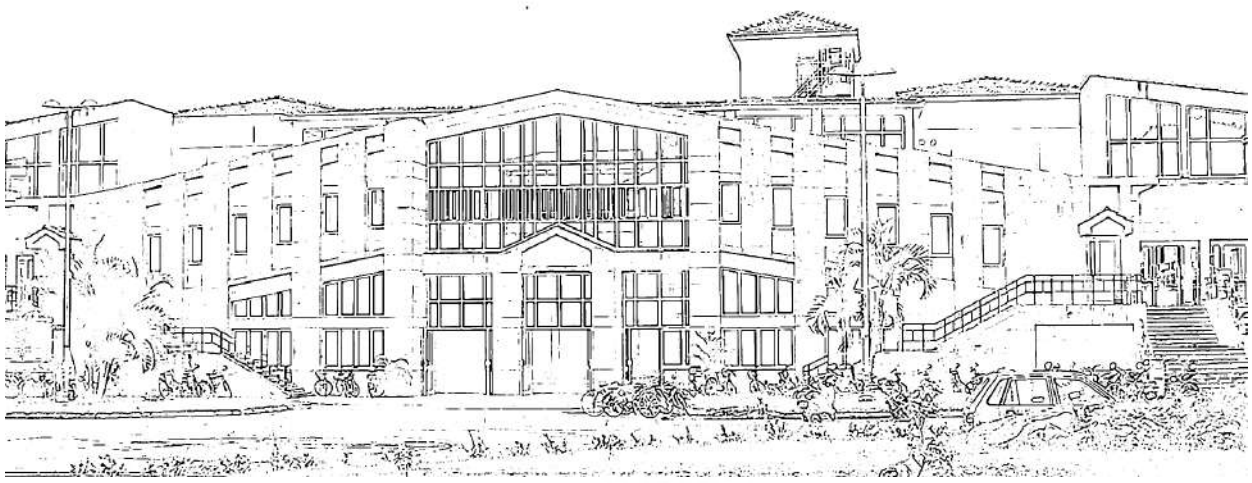
[199] J. Zhang, J. Huang, M. Yao, Z. Yang, H. Yu, M. Zhou, and F. Zhao. "Ingredient-Oriented Multi-Degradation Learning for Image Restoration." In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5825–5835. 2023.

[200] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu. "Segnext: Rethinking Convolutional Attention Design for Semantic Segmentation." *Advances in Neural Information Processing Systems*. vol. 35. pp. 1140–1156. 2022.

[201] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu. "LatticeNet: Towards Lightweight Image Super-Resolution with Lattice Block." In Proc. of the European Conference on Computer Vision (ECCV). pp. 272–289. 2020.

[202] Z. Chen, C. Liu, K. Zhang, Y. Chen, R. Wang, and X. Shi. "Underwater Image Super-Resolution via Range-Dependency Learning of Multiscale Features." *Computers and Electrical Engineering*. vol. 110. p. 108756. 2023.

[203] A. Pramanick, D. Megha, and A. Sur. "Attention-Based Spatial-Frequency Information Network for Underwater Single Image Super-Resolution." In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3560–3564. 2024.



Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati 781039, India