MSTR-Net: A Multi-Scale Transformer Refinement Network for Underwater Image Enhancement

Pranjali Singh, *Student Member, IEEE,* and Prithwijit Guha, *Senior Member, IEEE*

***Abstract*—Underwater images often suffer from severe visual degradation caused by light scattering, absorption, and wavelength-dependent attenuation, leading to reduced visibility, color distortion, and loss of structural details. To address these challenges, this paper proposes *MSTR-Net*, a *Multi-Scale Transformer Refinement Network* for underwater image enhancement (UIE). The proposed framework adopts a transformer-based encoder with efficient self-attention to capture long-range contextual information while maintaining computational efficiency. Hierarchical features extracted at multiple spatial resolutions are aggregated through a multi-scale feature fusion decoder, which predicts a residual image to restore fine structural details. To further improve visual quality, a multi-scale structural refinement module enhances edge continuity and suppresses local artifacts, while a luminance-guided refinement module corrects uneven illumination without introducing color distortion. As a result, *MSTR-Net* produces visually consistent outputs with improved color balance and sharper boundaries, effectively avoiding over-saturation and halo artifacts commonly observed in existing methods. Extensive experiments conducted on four benchmark underwater image datasets demonstrate that the proposed method consistently outperforms state-of-the-art approaches in both full-reference and no-reference evaluation metrics, confirming its effectiveness and robustness across diverse underwater conditions. The codes of *MSTR-Net* are available at https://github.com/pranjali1996/MSTR-Net.**

***Index Terms*—Efficient attention, Feature refinement, Hybrid attention, Multi-stage fusion, Underwater Image Enhancement.**

I. Introduction

Underwater images often suffer from severe degradation caused by light absorption, scattering, and wavelength-dependent attenuation. These effects reduce visibility, distort colors, and degrade structural details, which negatively impacts both visual quality and downstream vision tasks [1].

Early UIE methods rely on physical imaging models and simplified assumptions about water conditions [2, 3]. While computationally efficient, such methods often fail in real-world scenarios with varying illumination and water properties, leading to unstable results and color artifacts [4].

Learning-based UIE methods have shown strong performance by learning enhancement mappings from data. CNN-based approaches, such as WaterNet and UCOLOR, improve color and contrast [5, 6], while GAN-based methods enhance visual appearance through adversarial learning [7, 8]. However, most learning-based approaches rely on local convolutional operations, which limits their ability to capture global context and long-range dependencies [9, 10].

Transformer-based architectures address this limitation by modeling long-range dependencies using self-attention [11]. Hierarchical designs, such as the Swin Transformer, improve efficiency for high-resolution images [12]. Nevertheless, standard self-attention is computationally expensive, and preserving fine structures and natural color remains challenging in UIE tasks [13].

To overcome these limitations, this paper proposes *MSTR-Net*, a *Multi-Scale Transformer Refinement Network* for UIE. MSTR-Net combines a hybrid transformer-based encoder with multi-scale feature fusion and refinement modules to jointly model global context and local structures. A luminance-guided refinement module further improves brightness and contrast while preserving natural colors. Experimental results on multiple benchmark datasets demonstrate that MSTR-Net consistently enhances underwater image quality under diverse conditions.

The main contributions of this work are summarized as follows. First, the *MSTR-Net* is proposed as a multi-scale transformer refinement network for UIE that jointly models global context and local structures through a hierarchical encoder–decoder framework. Second, a *hybrid-attention-based multi-scale encoder (LoG-Net)* is designed, combining local convolutional encoding in shallow stages with global self-attention in deeper stages to balance contextual modeling and computational efficiency. Also, a *Local Enhancement Block (LEB)* is integrated within each encoder block to strengthen neighborhood-level feature representation and improve texture continuity. Third, A *High-Frequency Boost (HFB)* module is introduced at the input of the encoder to explicitly enhance edge sharpness and texture information suppressed by underwater scattering effects. Fourth, A *Multi-Scale Structural Refinement (MSR)* module is incorporated in the decoder to improve spatial consistency and suppress local enhancement artifacts across different receptive fields, along with a *Luminance-Guided Refinement (YR)* module, which is introduced to enhance brightness and contrast while preserving natural color relationships. The *MSTR-Net* is evaluated on a *synthetically degraded underwater image dataset ($UIEB-D8$ and $EUVP-X-D8$)* [14], consisting of eight degradation types and one clear class. Extensive experiments are conducted on *four benchmark underwater image datasets*, demonstrating consistent qualitative and quantitative improvements achieved by MSTR-Net.

Pranjali Singh is with the Centre for Intelligent Cyber-Physical Systems, Indian Institute of Technology Guwahati, Assam, India (e-mail: s.pranjali@iitg.ac.in).

Prithwijit Guha is with the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Assam, India (e-mail: pguha@iitg.ac.in).

II. Related Work

Existing approaches to Underwater Image Enhancement (UIE) can be broadly grouped into two categories – (a) Physics based methods, and (b) Learning based techniques. The later is further extended with Transformer based models.

Physics-Based UIE methods aim to reverse underwater degradation using simplified imaging models based on light absorption and scattering properties [1, 15]. Early approaches estimate transmission maps or background light to restore visibility [2], while dehazing-based techniques adapt atmospheric scattering models to underwater environments [3]. Although these methods are interpretable and computationally efficient, they rely on strong assumptions about water conditions, which often break down in real-world scenarios, leading to unstable results and color artifacts [4].

Learning-Based UIE approaches have been proposed to enhance underwater images using deep neural networks. WaterNet employs a fusion-based strategy to combine multiple enhancement priors, resulting in improved color and contrast correction [5]. UCOLOR extends this approach by leveraging multi-color space representations guided by underwater imaging characteristics, leading to better color fidelity [6]. GAN-based approaches, such as FUnIE-GAN and UGAN, use adversarial learning to enhance visual appearance [7, 8]. However, these methods often suffer from training instability and may introduce perceptual artifacts. More recent methods, including SGUIE-Net and HCLR-Net, incorporate semantic attention and contrastive learning to improve robustness and generalization under diverse underwater conditions [9, 10]. Semi-UIR further explores semi-supervised learning to utilize unlabeled data for UIE [16]. Despite these advances, most learning-based approaches rely primarily on local convolutional operations, which limits their ability to capture long-range dependencies.

Transformer based image enhancement methods have been introduced to image enhancement tasks due to their strong ability to model long-range dependencies via self-attention mechanisms [11]. Hierarchical vision transformers, such as the Swin Transformer, improve scalability and computational efficiency for high-resolution images [12]. However, standard self-attention exhibits quadratic computational complexity, making it less suitable for dense prediction tasks such as UIE. Recent studies therefore investigate efficient attention mechanisms and hybrid convolution–transformer designs to balance global context modeling and efficiency [13]. Nonetheless, preserving fine textures and natural color appearance remains challenging in transformer-based UIE methods.

A. Synthetic Underwater Image Datasets

Large-scale datasets are essential for training robust UIE models; however, collecting paired real underwater images remains costly and challenging. To address this issue, several works generate synthetic underwater images to simulate common degradation effects [8, 17]. While synthetic data improves diversity, it may introduce domain gaps when applied to real-world images [18]. Consequently, combining synthetic and real datasets is widely adopted to improve generalization. Following this strategy, the $UIEB-D8$ and $EUVP-X-D8$ datasets are used in this work. The dataset construction protocol and degradation modeling can be found in [14], and are thus not repeated here. $UIEB-D8$ and $EUVP-X-D8$ [14] enable controlled evaluation under diverse degradation conditions and hence complement the real-world benchmark datasets.

In summary, existing UIE methods span physics-based, learning-based, and transformer-based approaches, each with distinct strengths and limitations. These observations motivate the development of efficient multi-scale architectures that jointly model global context and local structures with dedicated refinement mechanisms, as pursued in this work.

III. Methodology

This section presents the proposed *MSTR-Net* framework for UIE. The architecture integrates global context modeling, local feature enhancement, and refinement-based restoration within a unified encoder–decoder paradigm. The complete block diagram of the proposed method is illustrated in Fig. 1.

A. Overall Architecture

Let $\mathbf{I}_d \in \mathbb{R}^{h\times w\times 3}$ denote a degraded underwater image and $\mathbf{I}_c \in \mathbb{R}^{h\times w\times 3}$ its corresponding reference image. The objective of MSTR-Net is to estimate an enhanced image $\mathbf{I}_e \in \mathbb{R}^{h\times w\times 3}$ that restores visibility, contrast, color balance, and structural details, addressing diverse underwater degradations. The framework follows an encoder–decoder architecture augmented with residual learning and successive refinement stages. The overall enhancement process is formulated as

$$\mathbf{I}_e = \mathcal{R}_Y(\mathcal{R}_{MS}(\mathbf{I}_d + \mathcal{D}(\mathcal{E}(\mathbf{I}_d)))) \tag{1}$$

where $\mathcal{E}(\cdot)$ denotes the encoder, $\mathcal{D}(\cdot)$ denotes the decoder, $\mathcal{R}_{MS}(\cdot)$ represents multi-scale structural refinement, and $\mathcal{R}_Y(\cdot)$ denotes luminance-guided refinement.

Let $\mathbf{T}_{out} = Conv(\mathbf{T}_{in}; k, s, c_{out}, g, \gamma_f)$ be the convolution operation on input tensor $\mathbf{T}_{in} \in \mathbb{R}^{h\times w\times c_{in}}$ by using c_{out} number of $k\times k\times c_{in}$ convolution kernels with stride s, padding $p = \frac{k-1}{2}$ and group-convolution parameter g ($g \geq 1$). Here, $g = 1$ and $g = c_{in}$ respectively imply no group- convolution and depth-wise convolution. The output tensor $\mathbf{T}_{out}$ is further subjected to the activation function γ_f. This work uses either $\gamma_f = ReLU$ or $\gamma_f = GELU$ or $\gamma_f = none$ (no activation).

B. High-Frequency Boost (HFB)

Underwater images often suffer from loss of fine details and edge sharpness due to scattering and absorption effects. To explicitly enhance high-frequency information, a *HFB* module is applied at the encoder input.

Given a degraded input image $\mathbf{I}_d \in \mathbb{R}^{h\times w\times 3}$, a low-frequency approximation is first obtained using a depthwise convolution:

$$\mathbf{I_{low}} = Conv(\mathbf{I}_d; 3, 1, c_{in}, c_{in}) \tag{2}$$

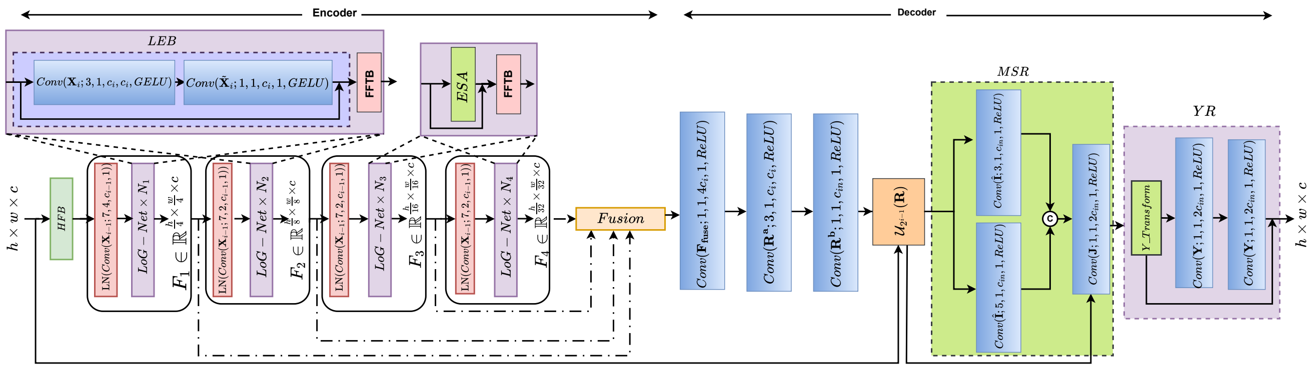

Fig. 1: Functional block diagram of the proposed MSTR-Net architecture.

The high-frequency component is then computed as the residual between the original image and its low-frequency approximation:

$$\mathbf{I}_{\mathbf{high}} = \mathbf{I}_d - \mathbf{I}_{\mathbf{low}} \tag{3}$$

The boosted representation is obtained by reinforcing the original input with the extracted high-frequency information:

$$\tilde{\mathbf{I}} = \mathbf{I}_d + \mathbf{I}_{\mathbf{high}} \tag{4}$$

All operations in the HFB module preserve the spatial resolution and channel dimensionality of the input, where $\mathbf{I}_{\mathbf{low}} \in \mathbb{R}^{h\times w\times c_{in}}$, $\mathbf{I}_{\mathbf{high}} \in \mathbb{R}^{h\times w\times c_{in}}$, and $\tilde{\mathbf{I}} \in \mathbb{R}^{h\times w\times c_{in}}$. By explicitly amplifying high-frequency details, the HFB module enhances edge sharpness and texture clarity at early stages of the network, providing a stronger foundation for subsequent feature extraction and attention-based processing.

C. Multi-Scale Transformer Encoder

The encoder extracts hierarchical feature representations from four $(N_f = 4)$ stages, with each stage progressively reducing spatial resolution while increasing semantic abstraction. This hierarchical design enables effective modeling of both local structures and global contextual information.

1) Patch Embedding (PE): Patch embedding is implemented using overlapped strided convolutions to progressively downsample spatial resolution while increasing channel dimensionality across encoder stages. This strategy preserves local continuity and mitigates blocking artifacts, making it suitable for dense image restoration tasks. Let $\mathbf{X}_0 \in \mathbb{R}^{h\times w\times c_0}$ be the feature map input to the encoder. The embedded feature map $\mathbf{X}_i \in \mathbb{R}^{h\times w\times c_i}$ at the i-th encoder stage is obtained as

$$\mathbf{X}_i = \mathrm{LN}(Conv(\mathbf{X}_{i-1}; 7, s, c_{i-1}, 1)) \tag{5}$$

where $\mathbf{X}_{i-1} \in \mathbb{R}^{h\times w\times c_{i-1}}$ and $\mathrm{LN}(\cdot)$ represents layer normalization. Here, the stride values for the four stages are set to $s \in \{4, 2, 2, 2\}$.

2) Hybrid Attention (HA) Mechanism: To balance representational power and computational efficiency, an *HA* mechanism is employed. In shallow stages, attention is realized using local convolution-based operators that restrict interactions to spatial neighborhoods. Accordingly, a *Local Enhancement Block* (LEB) is integrated within each encoder stage to enhance fine-grained texture information. The LEB operation on $\mathbf{X}_i$ at the i^{th} encoder stage produces $\mathbf{O}_i \in \mathbb{R}^{h\times w\times c_i}$ in the following manner

$$\tilde{\mathbf{X}}_i = Conv(\mathbf{X}_i; 3, 1, c_i, c_i, GELU) \tag{6}$$

$$\mathbf{O}_i = \mathbf{X}_i + Conv(\tilde{\mathbf{X}}_i; 1, 1, c_i, 1, GELU) \tag{7}$$

The *efficient self-attention (ESA)* mechanism [19] decouples spatial interactions from token-wise dependencies. This enables global context modeling with reduced computation cost. Conventional self-attention incurs quadratic complexity with respect to the number of spatial tokens N. In contrast, ESA reduces the complexity to $\mathcal{O}(c_i^2)$, where $c_i \ll N$. In the proposed architecture, ESA is applied only in deeper encoder stages to capture long-range dependencies while maintaining overall efficiency.

3) Feed-Forward Transformation Block (FFTB): Each encoder block includes a convolutional feed-forward transformation block to enhance nonlinear representation capacity while preserving spatial correspondence. For the input feature map $\mathbf{X}_i \in \mathbb{R}^{h\times w\times c_i}$, the FFTB output $\mathbf{Z}_i \in \mathbb{R}^{h\times w\times c_i}$ is obtained as follows

$$\mathbf{Z}_i^a = Conv(\mathbf{O}_i; 1, 1, c_i, 1, none) \tag{8}$$

$$\mathbf{Z}_i^b = Conv(\mathbf{Z}_i^a; 3, 1, c_i, c_I, none) \tag{9}$$

$$\mathbf{Z}_i^c = Conv(\mathbf{Z}_i^b; 1, 1, c_i, 1, GELU) \tag{10}$$

$$\mathbf{Z}_i = Conv(\mathbf{Z}_i^c; 1, 1, c_i, 1, none) \tag{11}$$

D. Multi-Scale Feature Fusion Decoder

The decoder aggregates hierarchical features from all encoder stages. Let $\mathbf{F}_i \in \mathbb{R}^{\left(\frac{h}{2^{i+1}}\right)\times\left(\frac{w}{2^{i+1}}\right)\times c_i}$ $(i = 1, \ldots N_f)$ denote encoder feature maps at different resolutions. Each feature map is projected and spatially aligned as

$$\tilde{\mathbf{F}}_i = \mathcal{U}_{2^{i-1}}(Conv(\mathbf{F}_i; 1, 1, c_i, g_i = 1, none)) \tag{12}$$

where $\mathcal{U}(\cdot)$ denotes bilinear up-sampling by a factor of 2^{i-1}. The aligned features are concatenated as:

$$\mathbf{F}_{\text{fuse}} = \text{Concat}(\tilde{\mathbf{F}}_1, \dots, \tilde{\mathbf{F}}_{N_f}) \tag{13}$$

The decoder predicts a residual image that maps the fused multi-scale features. It is defined as

$$\mathbf{R}^{\mathbf{a}} = Conv(\mathbf{F}_{\mathbf{fuse}}; 1, 1, 4c_i, 1, ReLU) \tag{14}$$

$$\mathbf{R}^{\mathbf{b}} = Conv(\mathbf{R}^{\mathbf{a}}; 3, 1, c_i, c_i, ReLU) \tag{15}$$

$$\mathbf{R} = Conv(\mathbf{R}^{\mathbf{b}}; 1, 1, c_{in}, 1, ReLU) \tag{16}$$

where $\mathbf{F}_{\mathbf{fuse}} \in \mathbb{R}^{h \times w \times 4c_i}$ denotes the concatenated multi-scale feature map and $\mathbf{R} \in \mathbb{R}^{h \times w \times c_{in}}$ is the residual feature map. The restored image is obtained via a skip connection as shown below:

$$\hat{\mathbf{I}} = \mathbf{I}_d + \mathbf{R} \tag{17}$$

where $\hat{\mathbf{I}} \in \mathbb{R}^{h \times w \times c_{in}}$ denotes the reconstructed output.

E. Multi-Scale Structural Refinement (MSR)

Although residual learning restores most visual content, structural inconsistencies and local artifacts may still persist in the intermediate output. To alleviate this issue, a *MSR* module is applied to the restored image obtained in (17). The *MSR* module explicitly enhances structural coherence by extracting complementary features using multiple receptive fields.

$$\mathbf{F}^{\mathbf{a}} = Conv(\hat{\mathbf{I}}; 3, 1, c_{in}, 1, ReLU) \tag{18}$$

$$\mathbf{F}^{\mathbf{b}} = Conv(\hat{\mathbf{I}}; 5, 1, c_{in}, 1, ReLU) \tag{19}$$

The refined output is obtained via residual fusion, given as:

$$\mathbf{J} = \text{Concat}(\ \mathbf{F}^{\mathbf{a}}\ ,\ \mathbf{F}^{\mathbf{b}}\ ) \tag{20}$$

$$\hat{\mathbf{J}} = Conv(\mathbf{J}; 1, 1, 2c_{in}, 1, ReLU) + \hat{\mathbf{I}} \tag{21}$$

Here, $\text{Concat}(\mathbf{F}^{\mathbf{a}}, \mathbf{F}^{\mathbf{b}}) \in \mathbb{R}^{h \times w \times 2c_{in}}$ denotes multi-scale structural feature maps extracted using different receptive fields. This is followed by convolution operation to fuse the multi-scale information before residual refinement.

F. Luminance-Guided Refinement (YR)

Luminance-Guided refinement or Y-Refinement *YR* module [5] is applied as a post-enhancement operation to explicitly address uneven illumination and reduced contrast. The luminance component $\mathbf{Y} \in \mathbb{R}^{h \times w \times 1}$ is extracted from $\hat{\mathbf{J}}$ using a standard linear transformation:

$$\mathbf{Y} = 0.299\,\hat{\mathbf{J}}_R + 0.587\,\hat{\mathbf{J}}_G + 0.114\,\hat{\mathbf{J}}_B \tag{22}$$

where $\hat{\mathbf{J}}_R$, $\hat{\mathbf{J}}_G$, and $\hat{\mathbf{J}}_B$ denote the red, green, and blue channels of $\hat{\mathbf{J}}$, respectively.

The extracted luminance is refined independently using a lightweight convolutional network:

$$\hat{\mathbf{Y}} = \mathcal{R}_Y(\ \mathbf{Y});\ \hat{\mathbf{Y}} \in \mathbb{R}^{h \times w \times 3} \tag{23}$$

where $\mathcal{R}_Y(\cdot)$ denotes a shallow refinement module composed of two convolutional layers with a $ReLU$ activation following each layer.

The refined luminance is then reintegrated into the RGB image through residual injection:

$$\mathbf{I}_e = \hat{\mathbf{J}} + \hat{\mathbf{Y}};\ \mathbf{I}_e \in \mathbb{R}^{h \times w \times 3} \tag{24}$$

By explicitly enhancing luminance information while leaving chromatic relationships intact, the *YR* module improves brightness and contrast without introducing color distortion, leading to visually natural restoration results.

IV. Experimental Setup

This section describes the datasets, training configuration, and evaluation metrics used to evaluate the proposed MSTR-Net.

A. Datasets

The proposed MSTR-Net is evaluated on four publicly available UIE benchmarks, namely *UIEB [5]*, *LSUI [20]*, *UFO-120 [21]*, and *EUVP [7]*. These datasets contain a diverse set of underwater scenes captured under varying water conditions, illumination levels, and degradation patterns. Together, they form a comprehensive benchmark for assessing the robustness and generalization performance of UIE methods. For all datasets, paired degraded and reference images are used for evaluation. Additionally, the synthetically degraded underwater image datasets $UIEB-D8$ and $EUVP-X-D8$ are used from [14] for training models. The $UIEB-D8$ and $EUVP-X-D8$ [14] datasets are divided into training and validation subsets using an 80:20 split. To maintain consistency across experiments, all images are resized to a fixed spatial resolution of 256×256 pixels. All images are converted to RGB format and normalized to the range $[0, 1]$. No additional data augmentation is applied unless otherwise stated.

B. Training Configuration

MSTR-Net is trained end-to-end on a Nvidia A100-80GB GPU using the mean squared error (MSE) loss between the enhanced output and the corresponding ground-truth image. The AdamW optimizer is used with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. The batch size is set to 8 for all experiments. The training is performed for a maximum of 1000 epochs. Early stopping is employed based on validation loss with a patience of 15 epochs to prevent overfitting. Gradient clipping with a maximum norm of 1.0 is applied to stabilize optimization. Mixed-precision training is adopted to improve computational efficiency.

TABLE I: Performance enhancement contributed by the inclusion of different modules in the proposed architecture.

Configuration					UIEB		EUVP		UFO-120		LSUI	
HA	LEB	HFB	MSR	YR	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
✓	×	×	×	×	23.7767	0.8724	22.8538	0.8316	25.3499	0.8336	22.5996	0.8714
✓	✓	×	×	×	24.4097	0.8834	23.7163	0.8374	26.0772	0.8381	23.4200	0.8728
✓	✓	✓	×	×	28.1532	0.9070	24.2260	0.8467	27.2346	0.8488	24.0531	0.8764
✓	✓	✓	✓	×	29.0013	0.9197	25.2882	0.8506	27.5072	0.8686	24.8256	0.8875
✓	✓	✓	✓	✓	**29.2323**	**0.9239**	**27.6852**	**0.9024**	**28.3233**	**0.8742**	**27.9704**	**0.9101**

C. Evaluation Metrics

The performance of MSTR-Net is evaluated using both full-reference and no-reference image quality metrics. For datasets with available ground-truth images, Peak Signal-to-Noise Ratio (PSNR) [22] and Structural Similarity Index Measure (SSIM) [23] are used to assess reconstruction fidelity. Higher PSNR and SSIM values indicate better enhancement performance. For real-world scenarios without reference images, Underwater Image Quality Measure (UIQM) [24] and Underwater Color Image Quality Evaluation (UCIQE) [25] are adopted to evaluate perceptual image quality. Higher values of UIQM and UCIQE correspond to improved visual quality.

PSNR and SSIM are reported for datasets with ground-truth references (UIEB [5], LSUI [20], UFO-120 [21]), while UIQM and UCIQE are used for real-world datasets without reference images (Challenge-60 [5], U45 [26], UCCS [27], and EUVP-330 [7].).

V. Results and Discussions

This section presents discussions on the ablation studies and experimental results, both qualitative and quantitative, associated with the performance analysis of the *MSTR-Net* system and its sub-systems.

A. Ablation Study on Architectural Modules

Table I reports the ablation results evaluating the contribution of individual modules across the UIEB [5], EUVP [7], UFO-120 [21], and LSUI [20] datasets using PSNR and SSIM metrics.

The baseline model with only Hybrid Attention (*HA*) provides moderate performance, capturing global context but lacking fine local detail restoration. Adding the Local Enhancement Block (*LEB*) consistently improves PSNR and SSIM on all datasets, highlighting the importance of localized feature refinement. Introducing the High Frequency Boost module (*HFB*) results in substantial gains, particularly in restoring edges and fine textures. Further improvements are achieved by incorporating the Multi-Scale Refinement module (*MSR*), which enhances structural consistency through multi-scale feature aggregation. The complete model, including the Luminance Refinement module (*YR*), achieves the best performance across all datasets. These results demonstrate the complementary contributions of each module and validate the effectiveness of the proposed architecture.

TABLE II: Performance comparison of different hybrid attention configurations on the UIEB [5] dataset.

Hybrid Attention Configuration	PSNR	SSIM
4 ESA	23.7935	0.8752
1 LEB + 3 ESA	28.0836	0.9152
2 LEB + 2 ESA	**29.2323**	**0.9239**
3 LEB + 1 ESA	26.4097	0.8834
4 LEB	23.4967	0.8634

TABLE III: Performance variations due to different refinement paths in the MSR module on the UIEB [5] dataset.

MSR Refinement Paths	PSNR	SSIM
$\mathrm{Conv}(\hat{\mathbf{I}}\ ;3\ ,1\ ,\mathrm{c}_{in},1,ReLU)$	27.1532	0.9070
$\mathrm{Conv}(\hat{\mathbf{I}}\ ;5\ ,1\ ,\mathrm{c}_{in},1,ReLU)$	28.2501	0.9050
$\mathrm{Conv}(\hat{\mathbf{I}}\ ;3\ ,1\ ,\mathrm{c}_{in},1,ReLU)+Conv(;5,1,c_{in},1,ReLU)$	**29.2323**	**0.9239**
$\mathrm{Conv}(\hat{\mathbf{I}}\ ;3\ ,1\ ,\mathrm{c}_{in},1,ReLU)+Conv(;5,1,c_{in},1,ReLU)+Conv(;7,1,c_{in},1,ReLU)$	27.6526	0.8812

B. Ablation Study on Hybrid Attention Configuration

Table II analyzes different configurations of the proposed hybrid attention mechanism on the UIEB [5] dataset. The hybrid design combines *LEB* with Efficient Self-Attention (*ESA*) to balance local detail restoration and global contextual modeling. Using only efficient attention (4 ESA blocks) yields limited performance, indicating insufficient recovery of fine-grained local structures. Replacing attention blocks with LEBs progressively improves performance up to an optimal point. The configuration with *2 LEB + 2 ESA* achieves the best results, attaining a PSNR of 29.23 dB and an SSIM of 0.9239.

When the number of LEBs exceeds the number of ESAs, performance degrades, suggesting that excessive local processing without adequate global context leads to suboptimal enhancement. Similarly, using only local information (4 LEBs) yields the weakest performance, underscoring the need for global attention for effective UIE. Overall, this study demonstrates that a balanced integration of LEB and ESA is critical and validates the finalized hybrid configuration adopted in the proposed architecture.

C. Effect of Refinement Paths in MSR

Table III analyzes the impact of different refinement paths within the MSR module on the UIEB [5] dataset.

Using a single refinement path with $Conv(\hat{\mathbf{I}};3,1,c_{in},1,ReLU)$ or $Conv(\hat{\mathbf{I}};5,1,c_{in},1,ReLU)$ kernels provides moderate improvements, capturing either fine or mid-level contextual information. The combination of

TABLE IV: Comprehensive quantitative comparison of *MSTR-Net* with state-of-the-art UIE methods. Full-reference metrics (PSNR/SSIM) are reported on UIEB [5], LSUI [20], and UFO-120 [21], while no-reference metrics (UIQM/UCIQE) are reported on Challenge-60 [5], U45 [26], UCCS [27], and EUVP-330 [7]. The best, second-best, and third-best performances are highlighted in *red*, *green*, and *blue*, respectively.

Method	**GFLOPs ↓**	**Params (M) ↓**	**Full-reference Metrics**						**No-reference Metrics**							
			UIEB		LSUI		UFO-120		Challenge-60		U45		UCCS		EUVP-330	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	UIQM	UCIQE	UIQM	UCIQE	UIQM	UCIQE	UIQM	UCIQE
WaterNet [5]	193.7	24.81	21.04	0.860	22.74	0.852	23.84	0.818	4.195	0.585	4.167	0.596	4.158	0.565	3.810	0.602
UCOLOR [6]	443.85	157.0	21.03	0.876	24.43	0.848	26.95	0.887	3.992	0.567	4.144	0.579	3.583	0.535	3.552	0.569
UWNet [28]	14.79	0.335	17.36	0.795	22.01	0.848	25.02	0.850	3.832	0.506	3.792	0.518	3.194	0.482	3.429	0.516
ADMNet [29]	10.98	0.079	20.77	0.879	26.11	0.850	26.67	0.875	4.190	0.569	4.247	0.583	4.062	0.547	3.248	0.571
SCNet [30]	94.22	1.0	21.37	0.889	27.47	0.891	27.34	0.860	4.259	0.581	4.456	0.609	4.386	0.572	3.857	0.590
STSC [31]	819.37	33.0	21.05	0.866	27.67	0.879	26.06	0.884	4.261	0.575	4.343	0.605	4.008	0.541	3.717	0.602
U-Shape [32]	5.96	23.0	21.30	0.893	27.76	0.911	28.00	0.880	4.063	0.560	4.120	0.591	3.819	0.536	3.525	0.575
UIE-WD [33]	102.76	14.0	20.10	0.861	26.81	0.840	27.48	0.875	4.102	0.565	4.164	0.600	3.711	0.526	3.557	0.581
SGUIE [9]	156.4	14.0	21.12	0.888	27.77	0.912	28.30	0.878	4.221	0.579	4.236	0.600	3.967	0.532	3.663	0.592
FUnIEGAN [7]	143.97	7.02	19.12	0.832	24.23	0.857	26.79	0.837	4.325	0.581	4.372	0.617	4.086	0.548	3.636	0.596
CECF [34]	67.69	10.89	21.82	0.886	27.86	0.911	26.75	0.872	4.127	0.572	4.354	0.601	4.286	0.553	3.681	0.590
UGAN [8]	143.0	37.64	21.37	0.889	26.93	0.890	27.61	0.876	4.480	0.580	4.501	0.590	4.391	0.565	3.990	0.615
HCLR-Net [10]	5651.99	4.87	22.09	0.897	27.94	0.916	28.85	0.881	4.100	0.596	4.360	0.603	3.895	0.631	3.860	0.634
UIE-DM [35]	48.40	10.0	23.03	0.909	27.94	0.915	27.61	0.882	3.874	0.563	4.097	0.597	4.168	0.579	3.817	0.603
GUPDM [36]	191.60	1.49	21.76	0.869	27.34	0.911	27.81	0.882	4.276	0.565	4.399	0.589	4.017	0.587	4.108	0.582
Semi-UIR [16]	550.76	2.0	22.79	0.908	25.94	0.889	27.36	0.878	4.098	0.576	4.431	0.599	4.019	0.599	3.660	0.597
UnFormer [37]	13.25	8.68	22.5	0.87	28.88	0.91	-	-	-	-	3.59	0.61	-	-	-	-
MSTR-Net (Ours)	27.74	105.20	29.23	0.9239	27.97	0.9101	28.92	0.8942	6.84	0.67	5.95	0.685	5.03	0.799	5.68	0.743

$Conv(\hat{\mathbf{I}}; 3, 1, c_{in}, 1, ReLU)$ *and* $Conv(\hat{\mathbf{I}}; 5, 1, c_{in}, 1, ReLU)$ *refinement paths* achieves the best performance, indicating that complementary local and contextual feature aggregation is critical for effective enhancement.

However, extending the refinement to include a $Conv(\hat{\mathbf{I}}; 7, 1, c_{in}, 1, ReLU)$ kernel leads to performance degradation, suggesting that excessive receptive field expansion introduces over-smoothing and weakens structural preservation. Moreover, increasing the receptive field increases the number of training parameters. These results validate the choice of a dual-path $Conv(\hat{\mathbf{I}}; 3, 1, c_{in}, 1, ReLU)$ *and* $Conv(\hat{\mathbf{I}}; 5, 1, c_{in}, 1, ReLU)$ MSR configuration in the finalized architecture.

D. Discussion of Results on State-of-the-Art Comparison

The proposed *MSTR-Net* is evaluated against state-of-the-art UIE methods using both full-reference and no-reference metrics, as summarized in Table IV. Full-reference evaluation is performed using PSNR and SSIM on the UIEB [5], LSUI [20], and UFO-120 [21] datasets, while no-reference perceptual quality is assessed using UIQM and UCIQE on the Challenge-60 [5], U45 [26], UCCS [27], and EUVP-330 [7] datasets.

MSTR-Net achieves the best performance on the UIEB [5] dataset, with *29.23 dB PSNR* and *0.9239 SSIM*, demonstrating superior fidelity and structural preservation. Competitive results are also obtained on LSUI [20] and UFO-120 [21], where *MSTR-Net* records *27.97 dB PSNR / 0.9001 SSIM* and *28.92 dB PSNR / 0.8942 SSIM*, respectively.

In terms of no-reference evaluation, *MSTR-Net* consistently outperforms competing methods across all datasets. Notably, it achieves a UIQM of *6.84* and a UCIQE of *0.67* on the Challenge-60 [5] dataset, with similarly strong results on U45 [26], UCCS [27], and EUVP-330 [7]. These gains reflect improved color balance, contrast, and visual clarity.

Despite its strong enhancement performance, *MSTR-Net* maintains a favorable efficiency profile with *27.74 GFLOPs*, offering a balanced trade-off between restoration quality and computational complexity. Overall, the results confirm the effectiveness of the proposed hybrid attention and multi-stage refinement framework for UIE.

TABLE V: Ablation study on the impact of encoder depth using alternative layer configurations.

Configuration	**Dataset**	**PSNR**	**SSIM**
Layers (2,2,6,2)	UIEB	24.41	0.8834
	UFO-120	26.58	0.8481
	LSUI	23.72	0.8728
Layers (3,3,12,3)	UIEB	24.05	0.8818
	UFO-120	25.97	0.8486
	LSUI	23.54	0.8719

TABLE VI: Comparison of FPS and Implementation Hardware among different UIE Models

Models	**FPS**	**PSNR**	**Implementation Hardware**
FUnIE-GAN [7]	120.37	21.53	NVIDIA GTX 1080 (TensorFlow)
Shallow-UWnet [28]	92.86	22.69	Not Reported
U-Shape [32]	15.24	21.30	NVIDIA RTX 3090 (PyTorch, Ubuntu 20)
PUIE [38]	45.52	21.382	NVIDIA RTX 2080Ti GPU
MSTR-Net (Ours)	109.24	29.23	Nvidia A100 GPU

E. Ablation Study on Number of Layers

This ablation study investigates the effect of encoder depth by evaluating alternative layer configurations while keeping all other architectural components fixed. As shown in Table V, reducing the number of layers leads to a noticeable drop in both PSNR and SSIM across all datasets, indicating insufficient capacity to model complex underwater degradations. In contrast, the proposed *MSTR-Net* adopts a deeper encoder configuration of *(3,4,18,3)*, which achieves the best quantitative performance and is therefore selected as the final architecture. The corresponding results are reported in the overall state-of-the-art comparison as shown in Table IV.

The average frames per second (FPS) of representative UIE models along with their implementation hardware are

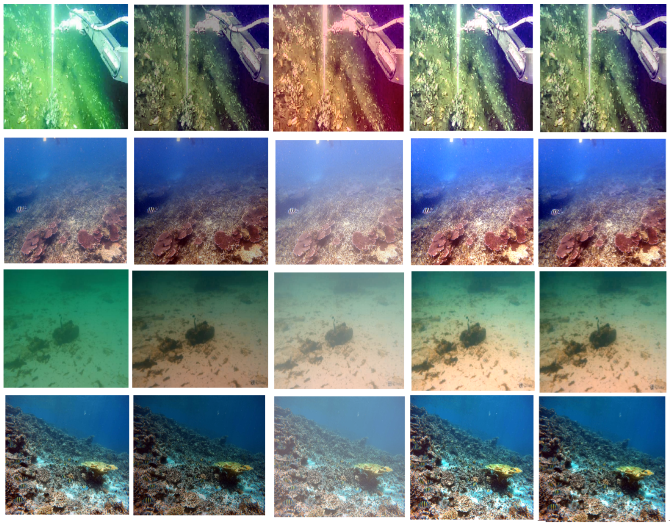

Fig. 2: Visual comparison of different enhancement methods on the UIEB [5] dataset, including the raw image, and results from *FUnIE-GAN [7]*, *HCLR-Net [10],* UGAN [8], and the proposed method.

reported in Table VI [1]. Lightweight models such as *FUnIE-GAN* and *Shallow-UWNet* achieve high FPS due to their shallow architectures, whereas deeper models like *U-Shape* exhibit lower runtime performance. The proposed *MSTR-Net* achieves a competitive FPS of *109.24* while delivering substantially higher enhancement quality (*29.23 dB PSNR*). These results demonstrate that *MSTR-Net* offers a favorable balance between computational efficiency and restoration performance compared to existing UIE methods.

F. Qualitative Performance Analysis

The qualitative performance of *MSTR-Net* is illustrated in Fig. 2. Sample images from the UIEB [5] test set are used to demonstrate the visual effectiveness of the proposed enhancement method. The qualitative results further support the quantitative findings, highlighting the superior performance of *MSTR-Net* compared to existing state-of-the-art approaches. *MSTR-Net* produces visually consistent results with improved color balance and sharper boundaries, avoiding over-saturation and halo artifacts observed in competing methods.

VI. Conclusion

This paper introduced *MSTR-Net*, a multi-stage refinement framework for UIE that effectively combines global contextual modeling, local feature enhancement, multi-scale, and luminance-guided refinements. Extensive evaluations on multiple benchmark datasets demonstrate that *MSTR-Net* consistently outperforms state-of-the-art methods in both full-reference (*PSNR by 6.2 dB*, *SSIM by 0.0159*) and no-reference (UIQM by 2.36, UCIQE by 0.168) metrics while maintaining moderate computational complexity. Overall, *MSTR-Net* offers an effective and practical solution for enhancing underwater imagery, with future work focusing on improving robustness under extreme degradations and extending the framework to video enhancement.

[1]The FPS comparison in Table VI involves significantly different hardware architectures, which limits the direct comparability of the runtime speeds.

References

[1] N. G. Jerlov, *Marine Optics*. Elsevier, 1976.

[2] T. Treibitz and Y. Y. Schechner, "Active polarization descattering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009.

[3] P. Drews, E. R. Nascimento, S. S. C. Botelho, and M. F. M. Campos, "Transmission estimation in underwater single images," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2016.

[4] S. Bazeille, I. Quidu, L. Jaulin, and J.-P. Malkasse, "Automatic underwater image preprocessing," in *CMM'06: Characterization, Measurement and Modeling of Natural Systems*, 2006.

[5] 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, IEEE.

[6] 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.

[7] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, 2020.

[8] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 7159–7165, IEEE.

[9] 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, IEEE.

[10] J. Zhou, J. Sun, C. Li, Q. Jiang, M. Zhou, K.-M. Lam, W. Zhang, and X. Fu, "HCLR-Net: Hybrid contrastive learning regularization with locally randomized perturbation for underwater image enhancement," *International Journal of Computer Vision*, vol. 132, no. 10, pp. 4132–4156, 2024, Springer.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words," *arXiv preprint arXiv:2010.11929*, vol. 7, p. 5, 2020.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[13] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, 2022.

[14] P. Singh and P. Guha, "Ida-uie: An iterative framework for deep network based degradation aware underwater image enhancement," in *Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing*. ACM, 2024, pp. 1–9.

[15] C. D. Mobley, *Light and Water: Radiative Transfer in Natural Waters*. Academic Press, 1994.

[16] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li, "Contrastive semi-supervised learning for underwater image restoration via reliable bank," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 145–18 155, IEEE.

[17] 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," in *IEEE Robotics and Automation Letters*. IEEE, 2017.

[18] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. IEEE, 2021, pp. 3531–3539.

[20] R. Pucci and N. Martinel, "CE-VAE: Capsule enhanced variational autoencoder for underwater image enhancement," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 2113–2123, IEEE.

[21] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," *arXiv preprint arXiv:2002.0115*, 2020.

[22] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.

[23] 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.

[24] 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.

[25] M. Yang and A. Sowmya, "Underwater image quality evaluation metrics," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6062–6071, 2015.

[26] 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.

[27] 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.

[28] A. Naik, A. Swarnakar, and K. Mittal, "Shallow-UWNet: Compressed model for underwater image enhancement (student abstract)," in *AAAI Conference on Artificial Intelligence*, vol. 35, no. 18, 2021, pp. 15 853–15 854, Proceedings of the AAAI.

[29] X. Yan, W. Qin, Y. Wang, G. Wang, and X. Fu, "Attention-guided dynamic multi-branch neural network for underwater image enhancement," *Knowledge-Based Systems*, vol. 258, p. 110041, 2022, Elsevier.

[30] Z. Fu, X. Lin, W. Wang, Y. Huang, and X. Ding, "Underwater image enhancement via learning water type desensitized representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2764–2768, IEEE.

[31] D. Wang, L. Ma, R. Liu, and X. Fan, "Semantic-aware texture-structure feature collaboration for underwater image enhancement," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 4592–4598, IEEE.

[32] 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, IEEE.

[33] Z. Ma and C. Oh, "A wavelet-based dual-stream network for underwater image enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2769–2773, IEEE.

[34] X. Cong, J. Gui, and J. Hou, "Underwater organism color fine-tuning via decomposition and guidance," in *AAAI conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1389–1398, Association for the Advancement of Artificial Intelligence.

[35] Y. Tang, H. Kawasaki, and T. Iwaguchi, "Underwater image enhancement by transformer-based diffusion model with non-uniform sampling for skip strategy," in *31st ACM International Conference on Multimedia*, 2023, pp. 5419–5427, ACM.

[36] P. Mu, H. Xu, Z. Liu, Z. Wang, S. Chan, and C. Bai, "A generalized physical-knowledge-guided dynamic model for underwater image enhancement," in *31st ACM International Conference on Multimedia*, 2023, pp. 7111–7120, ACM.

[37] Y. Qing, Y. Wang, H. Yan, X. Xie, and Z. Wu, "Unformer: A transformer-based approach for adaptive multi-scale feature aggregation in underwater image enhancement," *IEEE Transactions on Artificial Intelligence*, 2024.

[38] Z. Fu, W. Wang, Y. Huang, X. Ding, and K.-K. Ma, "Uncertainty inspired underwater image enhancement," in *European conference on computer vision*, 2022, pp. 465–482, Springer.