FUSION: Frequency-guided Underwater Spatial Image recOnstructioN

Supplementary Material

Additional Extended Methodology

In this section, we expand upon the mathematical foundations of our framework, detailing the operations performed in both the spatial and frequency domains, as well as their fusion and calibration.

1. Spatial Domain Processing: For an input image $D^{h\times w\times 3}$, each color channel D_i (with $i \in \{R, G, B\}$) is processed independently. The initial multi-scale feature extraction is given by:

$$f_i^1 = \Phi_i(D_i), \tag{18}$$

where $\Phi_i(\cdot)$ denotes convolutional operations with kernel sizes 3×3 (for R), 5×5 (for G), and 7×7 (for B). To enhance these features, a two-stage attention mechanism is applied.

First, channel attention is computed as:

$$W_{\text{channel}} = \sigma\Big(\mathbf{W}_2 \cdot \phi\big(\mathbf{W}_1 \cdot g(f_i^1)\big)\Big), \tag{19}$$

where $g(f_i^1)$ denotes global average pooling, $\phi(\cdot)$ is a ReLU activation, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices. The feature map is then scaled element-wise:

$$f_{\text{channel-att}} = W_{\text{channel}} \odot f_i^1. \tag{20}$$

Next, spatial attention is defined by:

$$W_{\text{spatial}} = \sigma\Big(\psi\big([\mathcal{P}_{avg}(f_{\text{channel-att}}); \mathcal{P}_{max}(f_{\text{channel-att}})]\big)\Big), \tag{21}$$

where $\mathcal{P}_{avg}$ and $\mathcal{P}_{max}$ denote average and max pooling, respectively, and $\psi(\cdot)$ is a convolutional mapping. The refined spatial features are obtained as:

$$f_i^2 = W_{\text{spatial}} \odot \big(W_{\text{channel}} \odot f_i^1\big). \tag{22}$$

Finally, a residual connection ensures low-level features are preserved:

$$f_i^3 = f_i^2 + f_i^1, \quad \forall i \in \{R, G, B\}. \tag{23}$$

2. Frequency Domain Processing: Each channel D_i is transformed into the frequency domain using the 2D Fast Fourier Transform (FFT):

$$F_i(u, v) = \sum_{x=0}^{h-1} \sum_{y=0}^{w-1} D_i(x, y)\, e^{-j2\pi\left(\frac{ux}{h} + \frac{vy}{w}\right)}. \tag{24}$$

The magnitude of the frequency representation is computed as:

$$|F_i(u, v)| = \sqrt{\text{Re}(F_i(u, v))^2 + \text{Im}(F_i(u, v))^2}. \tag{25}$$

To refine the magnitude features, we perform a linear transformation:

$$\hat{F}_i = W_2 \cdot \phi\Big(W_1 \cdot |F_i|\Big), \tag{26}$$

followed by frequency attention:

$$W_{\text{freq}} = \sigma\Big(W_4 \cdot \phi\big(W_3 \cdot \bar{F}_i\big)\Big), \quad \bar{F}_i = \frac{1}{hw} \sum_{u,v} |F_i(u, v)|. \tag{27}$$

The refined magnitude is:

$$|F_i|_{\text{refined}} = W_{\text{freq}} \odot |F_i|. \tag{28}$$

The phase information $\Theta_i(u, v)$ is preserved as:

$$\Theta_i(u, v) = \arctan\Big(\frac{\text{Im}(F_i(u, v))}{\text{Re}(F_i(u, v))}\Big), \tag{29}$$

and the refined complex representation is reconstructed by:

$$F_i'(u, v) = |F_i|_{\text{refined}} \cdot e^{j\Theta_i(u,v)}. \tag{30}$$

Finally, the inverse FFT recovers the spatial features:

$$f_{\text{freq},i}(x, y) = \frac{1}{hw} \sum_{u=0}^{h-1} \sum_{v=0}^{w-1} F_i'(u, v)\, e^{j2\pi\left(\frac{ux}{h} + \frac{vy}{w}\right)}. \tag{31}$$

3. Frequency Guided Fusion (FGF): The spatial features f_i^3 and frequency features $f_{\text{freq},i}$ are fused to form a unified representation:

$$f_{\text{concat},i} = \text{Concat}\Big(f_i^3,\, f_{\text{freq},i}\Big). \tag{32}$$

A convolutional layer then integrates these features:

$$f_{\text{fused},i} = \phi\Big(W_i * f_{\text{concat},i} + b_i\Big), \tag{33}$$

where $*$ denotes convolution and b_i is the bias term.

4. Inter-Channel Fusion and Channel Calibration: Fused representations from the three channels are concatenated:

$$f_{\text{all}} = \text{Concat}\Big(f_{\text{fused},R},\, f_{\text{fused},G},\, f_{\text{fused},B}\Big). \tag{34}$$

This aggregated feature is projected into a higher-dimensional space:

$$f_d = \phi\Big(\mathcal{T}_d(f_{\text{all}})\Big), \tag{35}$$

and further integrated with frequency features through a learned transformation:

$$f_{\text{fusion}} = \phi\Big(\mathcal{T}_f(f_d, f_{\text{freq}})\Big). \tag{36}$$

A global attention mechanism refines this fused representation:

$$f_{\text{attn}} = \mathcal{A}\Big(f_{\text{fusion}},\ f_{\text{all}}\Big), \tag{37}$$

followed by the reconstruction of a preliminary enhanced image:

$$E = \phi\Big(\mathcal{T}_e(f_{\text{attn}})\Big). \tag{38}$$

Finally, adaptive channel calibration is performed:

$$W_{\text{calibration}} = \sigma\Big(W_2 \cdot \phi\big(W_1 \cdot g(E)\big)\Big), \tag{39}$$

$$E_{\text{final}} = E \odot W_{\text{calibration}}, \tag{40}$$

ensuring that the final enhanced image E_{final} exhibits balanced color distributions and preserved structural details.

Hardware and Training Details

We run all our experiments on a NVIDIA Tesla P100 GPU (Pascal architecture) with 16 GB of HBM2 memory and 3,584 CUDA cores, delivering up to 9.3 TFLOPS of single-precision performance. Since our method focuses on lightweight design and real-time feasibility, testing on a GPU with minimal compute ensures efficiency without relying on heavy hardware. We also use automatic mixed-precision (AMP) to speed up training and reduce memory usage, making the process even more efficient.

Training Settings: We train our model using the Adam optimizer with a starting learning rate of 2×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. Training runs for up to 1,000 epochs with a batch size of 4, but we use early stopping based on LPIPS. We choose LPIPS since it closely aligns with human perception, ensuring that the model focuses on producing visually improved underwater images.

Computation-Related Info:

```
===== Model Performance Report =====
GPU Memory Used: 50.46 MB
Peak GPU Memory: 260.34 MB
Inference Time: 1.8147 seconds
Estimated FPS: 0.55 frames per second
Total FLOPs: 18.41 GFLOPs
```

Supplementary Results

+

In these supplementary results, we provide additional quantitative and qualitative visualizations to further illustrate the performance and efficiency of our proposed FUSION framework. In addition to the primary metrics presented in the main paper, these supplementary results include detailed ablation studies, bar plots comparing quality metrics across the UIEB, EUVP, and SUIM-E datasets,

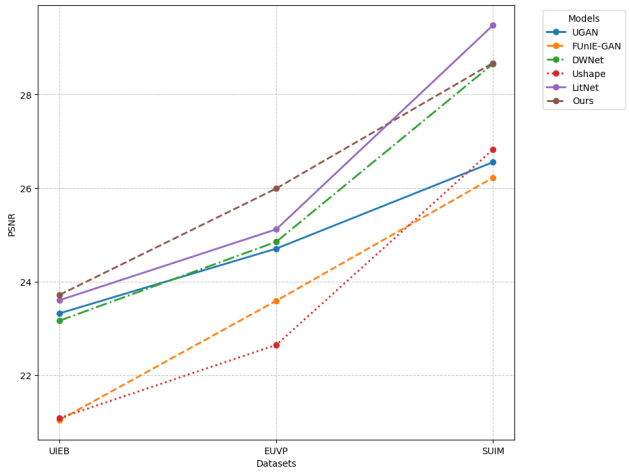


Figure 7. Line chart comparing PSNR values across the UIEB, EUVP, and SUIM-E datasets.

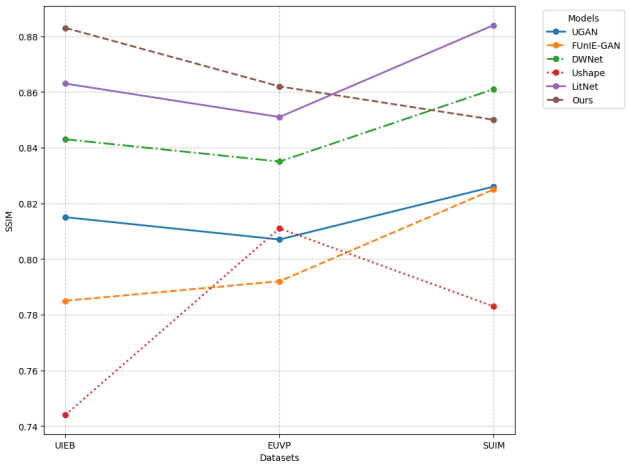


Figure 8. Line chart comparing SSIM values across the UIEB, EUVP, and SUIM-E datasets.

as well as extended efficiency analyses. These visualizations are closely tied to the mathematical formulations described in Section 3 and underscore the importance of our dual-domain processing.

The above figure (9) illustrates the impact of ablating key components of our proposed model. We observe that removing individual modules results in visible degradations, which verifies the necessity of each part for achieving optimal performance.

Figure 10 presents a qualitative comparison between our FUSION framework and several traditional image processing techniques. Notably, while methods such as histogram equalization and dark channel priors provide some level of enhancement, they fall short of recovering natural color balance and structural details, as achieved by our method.

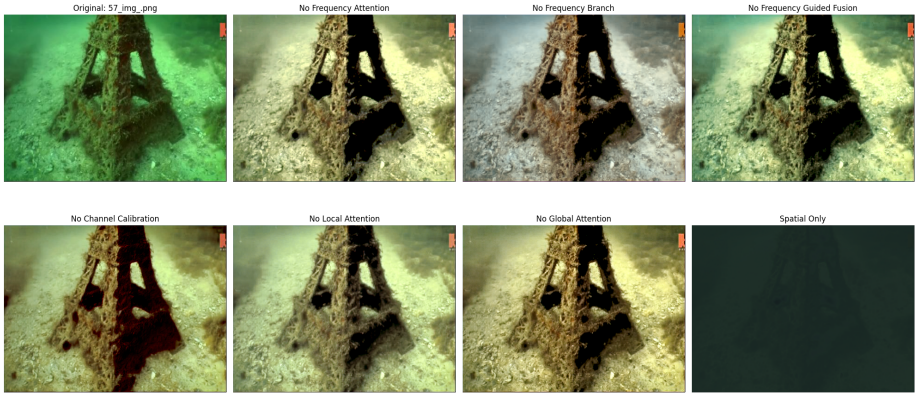


Figure 9. Ablation Study Visual Comparisons. This figure displays the enhancement results for a representative underwater image using our model with various component ablations: Original, No Frequency Attention, No Frequency Branch, No Frequency Guided Fusion, No Channel Calibration, No Local Attention, No Global Attention, and Spatial Only. The qualitative differences underscore the contribution of each module to the final enhancement quality.

Table 7. Evaluation 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	LPIPS↓	UIQM	UISM	BRISQUE↓
UDCP [5]	12.074	0.513	0.270	1.648	7.537	22.788
GBdehaze [13]	14.339	0.599	0.355	2.255	7.400	20.175
IBLA [26]	18.024	0.685	0.209	1.826	7.341	20.957
ULAP [30]	19.148	0.744	0.231	2.115	7.475	21.250
CBF [2]	20.395	0.834	0.194	3.003	7.360	21.115
UGAN [6]	24.704	0.826	0.190	2.894	7.175	20.288
UGAN-P [6]	25.050	0.827	0.188	2.901	7.184	18.768
FUnIE-GAN [10]	23.590	0.825	0.189	2.918	7.121	22.560
SGUIE-Net [28]	25.987	0.857	0.153	2.637	7.090	25.927
DWNet [29]	24.850	0.861	0.133	2.707	7.381	20.757
Ushape [24]	22.647	0.783	0.213	2.873	7.061	22.876
Lit-Net [27]	25.117	0.884	0.118	2.918	7.368	19.602
FUSION (Ours)	25.989	0.850	0.118	3.183	7.679	18.655

Figures 7 and 8 plot PSNR and SSIM values across the datasets. The PSNR chart shows FUSION consistently achieves higher reconstruction fidelity with elevated PSNR values. The SSIM chart reveals superior structural similarity compared to other approaches, even under challenging conditions. These plots highlight that FUSION enhances local details and color balance while preserving global image structure, reinforcing its effectiveness in underwater image enhancement tasks.

On the SUIM-E test set (Table 7), our approach further confirms its robustness by achieving comparable PSNR, SSIM, and LPIPS scores. Additionally, FUSION exhibits favorable performance in perceptual quality metrics, with competitive UIQM, UISM, and BRISQUE scores across all datasets.

Our methodology leverages multi-scale convolutions, adaptive attention mechanisms, and frequency-domain transformations to address the complex degradations in underwater images. To offer deeper insight into our approach, we now provide additional mathematical details that further elaborate on the operations used in FUSION.

Metric-wise Bar Plots

To provide a granular view of the performance across different metrics, we present bar plots (Figures 13-18)for each quality measure across the UIEB, EUVP, and SUIM-E datasets. These plots allow us to compare how various methods perform in terms of perceptual quality (BRISQUE and LPIPS), reconstruction fidelity (PSNR and SSIM), reconstruction (MSE), and overall image quality (UIQM and UISM).

Additional Efficiency Analysis

The efficiency of underwater image enhancement models is crucial for practical deployment, particularly on autonomous underwater vehicles (AUVs) and other resource-constrained platforms. In this section, we provide a side-by-side comparison of the model parameters and computational complexity (GFLOPs) for various SOTA methods. As described in our methodology, the efficient design of FUSION is achieved by leveraging dual-domain processing and optimized fusion strategies, which are mathematically formulated in Equations (1) through (9) for the spatial and frequency domains, respectively.

Figure 12 illustrates the trade-off between the parameter count and GFLOPs. This side-by-side visualization clearly shows that FUSION achieves competitive computational efficiency, with a remarkably low parameter count (0.28M) while maintaining a GFLOPs score of 36.73G. This balance

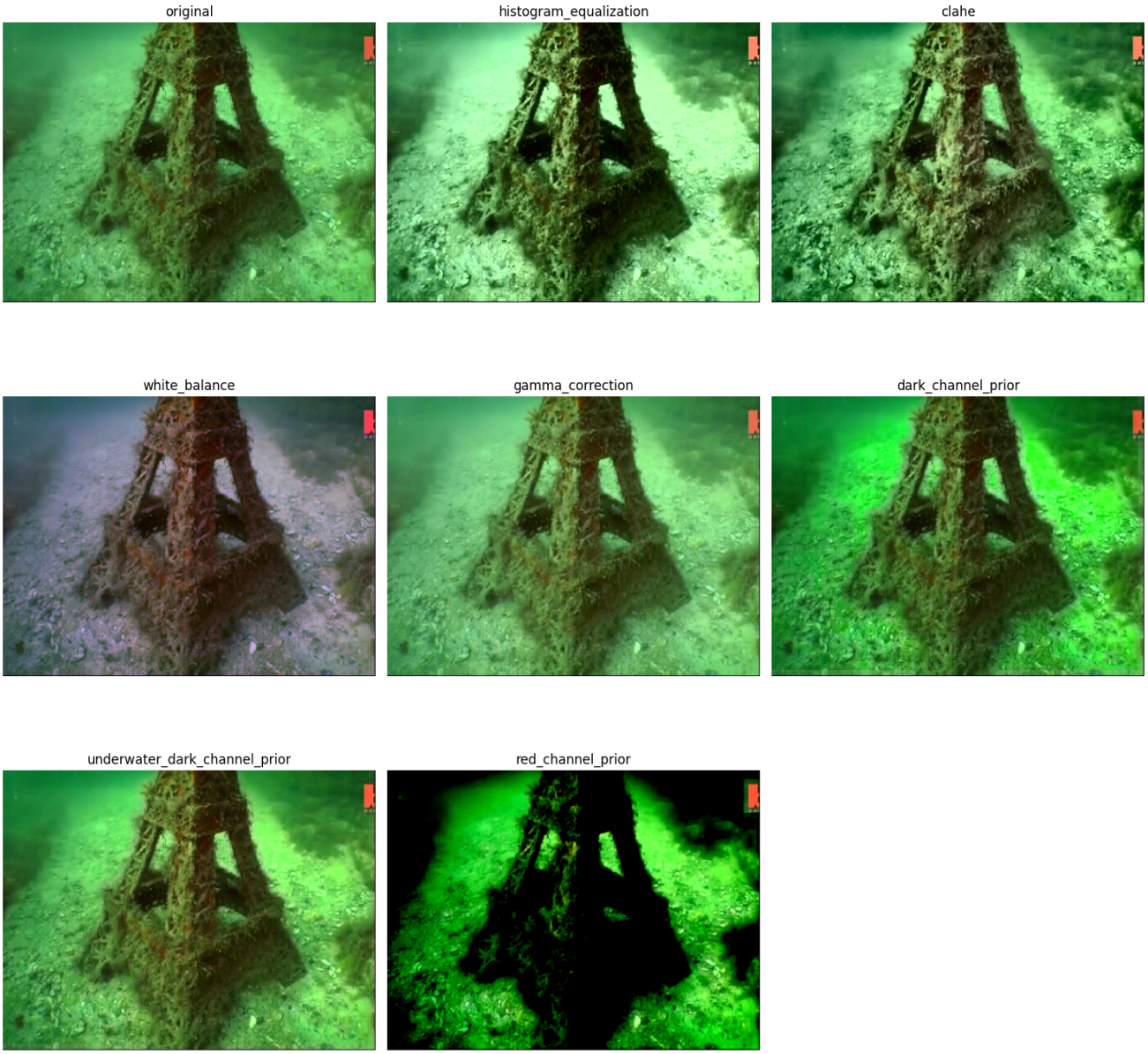


Figure 10. Comparison with Traditional Image Processing Techniques. The figure compares the original underwater image with images processed by conventional methods: histogram equalization, CLAHE, white balance, gamma correction, dark channel prior, underwater dark channel prior, and red channel prior. These comparisons highlight the limitations of traditional methods relative to our approach.

is a direct result of the adaptive attention mechanisms and efficient convolutional designs implemented within the network.

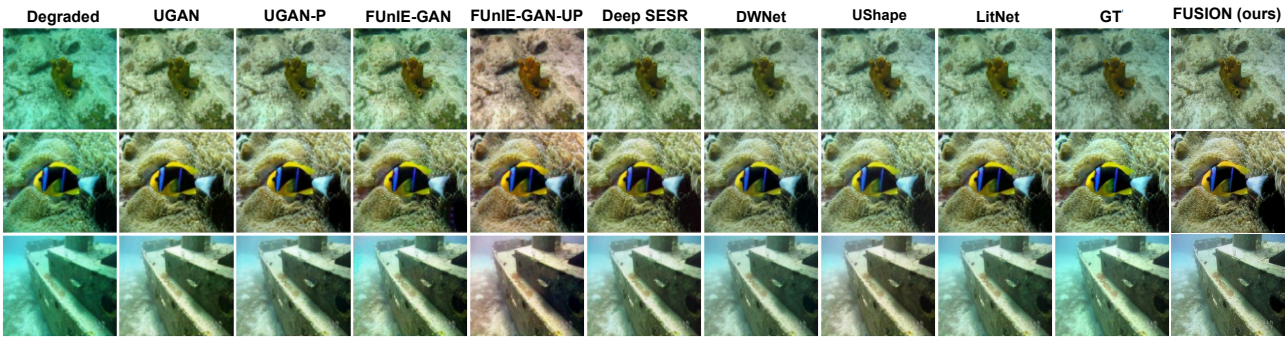


Figure 11. Visual comparisons on the SUIM-E dataset.

Table 8. Ablation performance on SUIM-E.

Config	Freq. Attn	Freq. Branch	Freq. Fusion	Chan. Calib	Local Attn	Global Attn	UIQM	UISM	LPIPS	BRISQUE
Full Model (FUSION)	✓	✓	✓	✓	✓	✓	3.183	7.679	0.118	18.655
no_frequency attention	✗	✓	✓	✓	✓	✓	2.626	5.832	0.242	23.91
no_frequency branch	✓	✗	✓	✓	✓	✓	2.806	5.674	0.285	25.73
no_frequency guided fusion	✓	✓	✗	✓	✓	✓	2.703	6.010	0.230	23.24
no_channel calibration	✓	✓	✓	✗	✓	✓	2.721	6.096	0.225	22.98
no_local attention	✓	✓	✓	✓	✗	✓	2.736	6.034	0.228	23.15
no_global attention	✓	✓	✓	✓	✓	✗	2.746	6.023	0.229	23.21
spatial only	✗	✗	✗	✓	✓	✓	2.645	5.506	0.265	24.62
minimal model	✗	✗	✗	✗	✗	✗	2.445	5.149	0.292	26.18

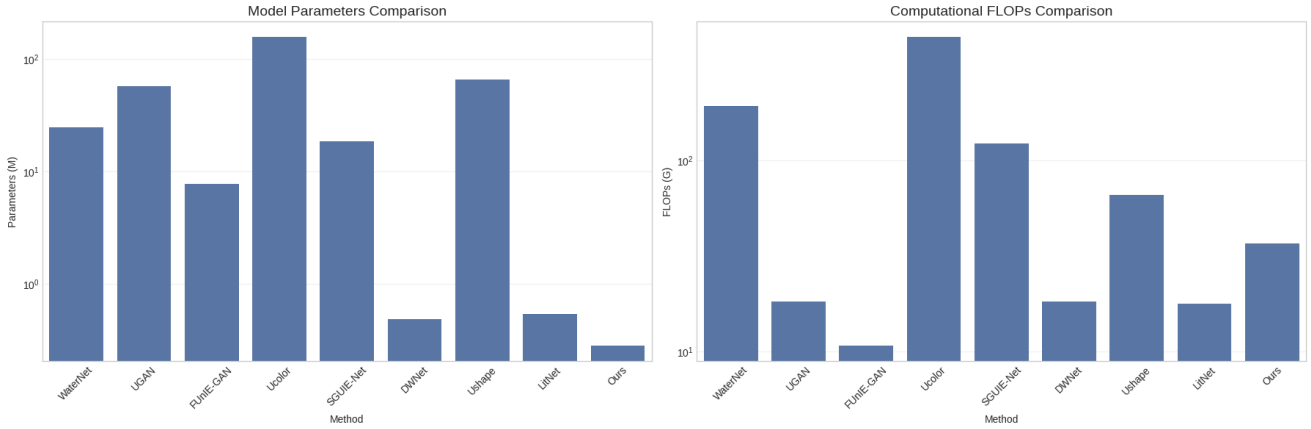
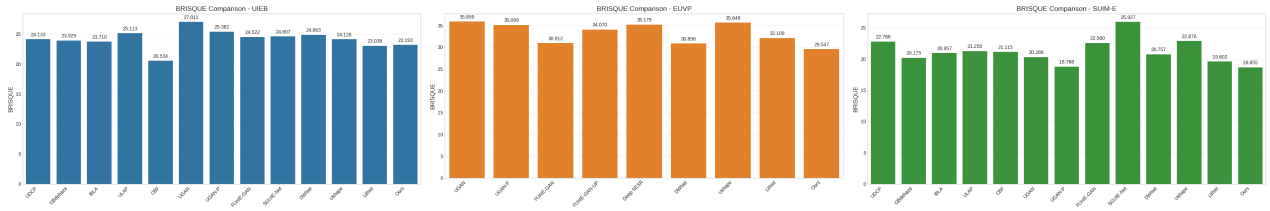
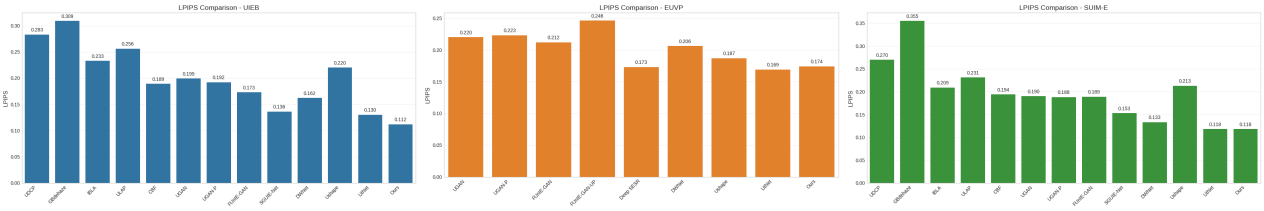


Figure 12. Side-by-side comparison of model parameters and GFLOPs for various UIE methods. FUSION achieves low computational cost without compromising enhancement performance.



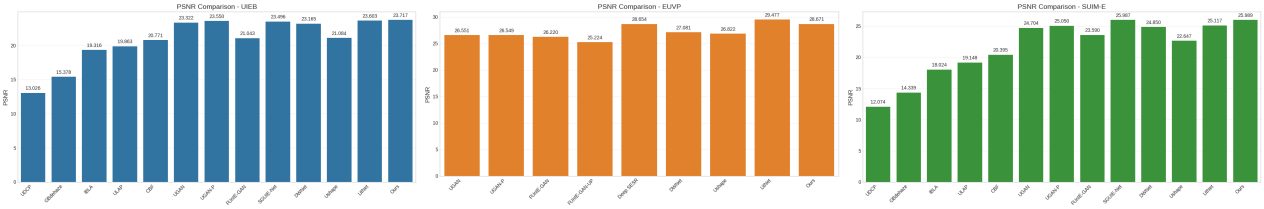
(a) UIEB (b) EUVP (c) SUIM-E

Figure 13. Bar plots comparing BRISQUE scores (lower is better) across the UIEB, EUVP, and SUIM-E datasets.



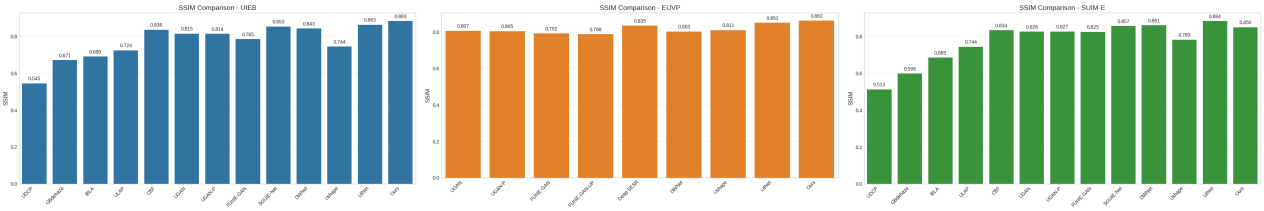
(a) UIEB (b) EUVP (c) SUIM-E

Figure 14. Bar plots comparing LPIPS scores (lower is better) across the three datasets. Lower LPIPS values indicate that FUSION produces enhanced images that are perceptually closer to the ground truth.



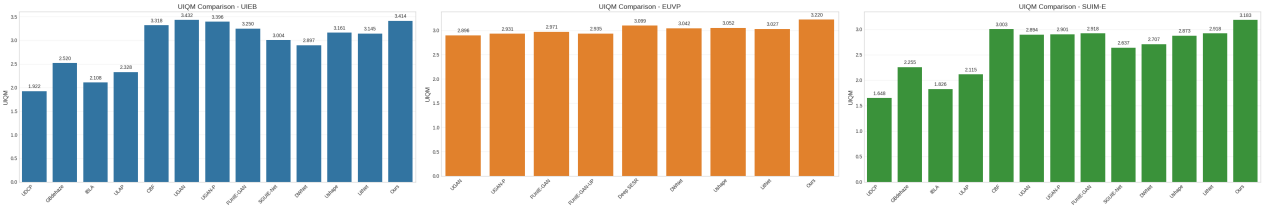
(a) UIEB (b) EUVP (c) SUIM-E

Figure 15. Bar plots comparing PSNR values across the UIEB, EUVP, and SUIM-E datasets. Higher PSNR values achieved by FUSION indicate its reconstruction fidelity.



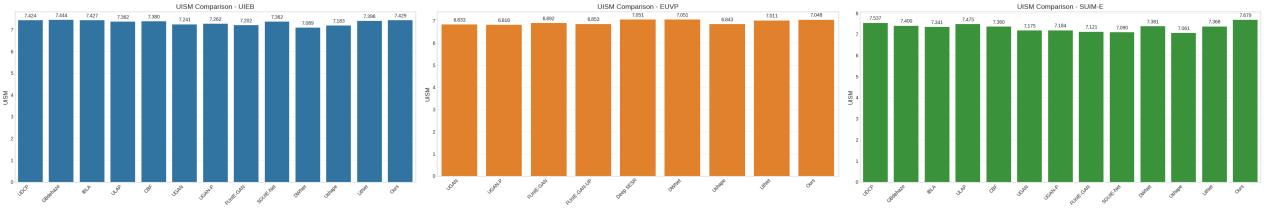
(a) UIEB (b) EUVP (c) SUIM-E

Figure 16. Bar plots comparing SSIM values across the three datasets. FUSION consistently achieves higher SSIM values.



(a) UIEB (b) EUVP (c) SUIM-E

Figure 17. Bar plots comparing UIQM scores across the UIEB, EUVP, and SUIM-E datasets. The UIQM metric reflects overall image quality improvements achieved by FUSION.



(a) UIEB (b) EUVP (c) SUIM-E

Figure 18. Bar plots comparing UISM scores across the three datasets. Higher UISM scores for FUSION indicate improved sharpness and detail retention in the enhanced images.