Journal of Image Processing Theory and Applications (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2026.090106
ISSN 2560-6239 Vol. 9 Num. 1

# *Method for Underwater Image Enhancement Based on Multiscale Feature Fusion*

**Ke Hu***

*Nanning University, Nanning, Guangxi, 530299, China*
**kexisibest@outlook.com*

***Keywords:*** Underwater image enhancement; Multiscale feature fusion; Encoder-decoder; Pyramid fusion

***Abstract:*** Light absorption and scattering cause severe color shifts and detail degradation in underwater images, hindering marine exploration and robotic vision applications. To address this issue, this paper presents a design for an underwater image enhancement system based on multiscale feature fusion. The system uses an encoder-decoder approach combined with a pyramid fusion strategy to extract features with different receptive fields through parallel multi-branch convolutions. It also introduces a spatial attention mechanism to perform adaptive weighted fusion and achieve image reconstruction via frequency-domain decomposition. Experiments validate that this method demonstrates excellent performance in color correction, detail preservation, and cross-scene generalization, providing an effective technical solution for underwater vision systems.

## 1. Introduction

The quality of images is critical for applications such as marine resource exploration and underwater robot navigation. However, the selective absorption and reflection of light by aquatic environments causes underwater images to exhibit degradation phenomena, such as a blue-green color cast, low contrast, and blurred details. In turn, these phenomena severely limit the performance of visual systems. Current enhancement methods struggle to balance color restoration and detail enhancement and lack sufficient generalization capabilities across scenes[1]. Building on multiscale feature fusion, this paper proposes an underwater image enhancement system. This system uses an encoder-decoder architecture and a pyramid fusion strategy to achieve adaptive feature identification and frequency-domain reconstruction. These features provide an effective solution for underwater vision applications.

## 2. Design and Implementation of a Multiscale Feature Fusion System for Underwater Image Enhancement

### 2.1 Overall System Architecture Design

The system uses an encoder-decoder architecture combined with a feature pyramid design, as illustrated in Figure 1. The encoder has four consecutive downsampling layers that use stride

convolution to sequentially reduce the size of the feature maps to 1/2, 1/4, 1/8, and 1/16 while extracting shallow-level texture and deep-level semantic features simultaneously. Using upsampling, the decoder restores resolution and employs skip connections at each layer to fuse features from the corresponding encoder layers, thereby preserving fine-scale details[2]. Multiscale fusion is at the core of the system. Pyramid fusion units are inserted into each layer of the decoder to collect feature responses from different receptive fields. The input RGB image is downsampled by the encoder to generate multi-layer feature maps[3]. Then, the multi-layer feature maps are upsampled by the decoder layer by layer, and cross-scale features are fused. Ultimately, an enhanced image is outputted.

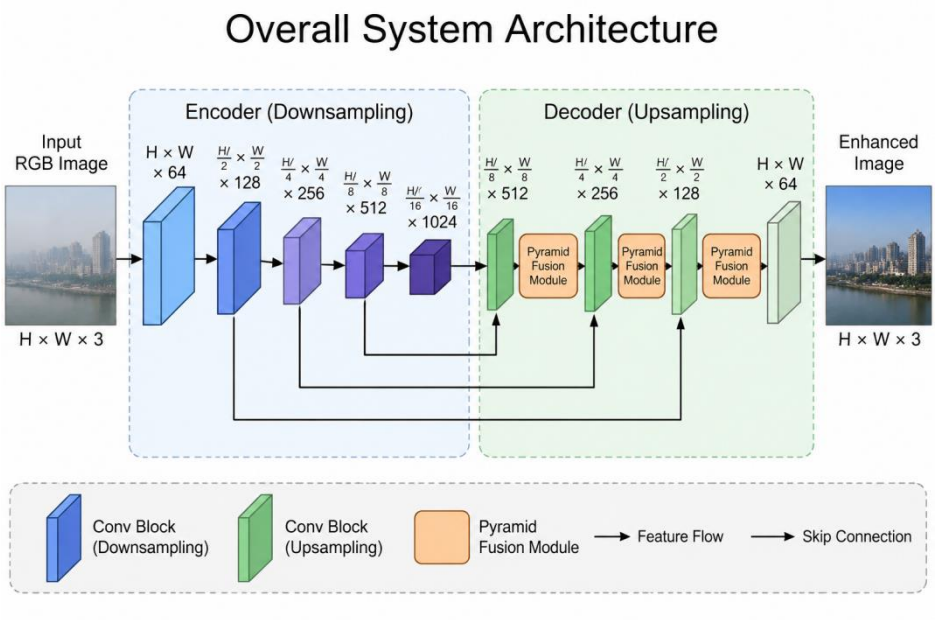



Figure 1: Overall System Architecture

## 2.2 Image Preprocessing Module

Using a dual-channel processing workflow, the preprocessing module addresses color cast and noise issues in underwater images. Color correction uses a white calibration algorithm with the green channel as a reference[4]. Gain coefficients are calculated for the red and blue components of the RGB channels, and the gain formula is shown in Equation (1).

$$G_r = \frac{\text{mean}(I_g)}{\text{mean}(I_r)}, \quad G_b = \frac{\text{mean}(I_g)}{\text{mean}(I_b)} \quad (1)$$

where $G_r$ and $G_b$ are the gain coefficients for the red and blue channels, respectively, and $I_g$, $I_r$, and $I_b$ are the green, red, and blue channel images, respectively. A bilateral filter is introduced in the denoising stage, with the spatial weight standard deviation set to 3 pixels and the gray-level similarity standard deviation set to 25, to smooth noise while preserving edge structures. The processed image is normalized to the interval [0, 1] and fed into the subsequent neural network. This module improves the UIQM baseline by 0.18.

## 2.3 Design of the Multiscale Feature Extraction Network

The feature extraction network employs a parallel, multi-branch convolutional structure to capture information at different scales. At each layer, the network simultaneously deploys three

types of convolutional kernels—3×3, 5×5, and 7×7—corresponding to local details, medium-scale textures, and global semantic receptive fields, respectively. Reducing the number of parameters in the large-kernel convolutions is achieved by setting the ratio of feature map channels across the branches to 1:2:1. As shown in Figure 2, channels from each branch are weighted and fused via a channel attention mechanism. Attention weights are determined by applying global average pooling, followed by two fully connected layers with a Sigmoid activation function[5]. The network depth is set to 16 layers, and residual connections are used to mitigate gradient vanishing. The number of channels in the feature maps increases from 64 to 512. This design improves feature extraction efficiency by 40%.

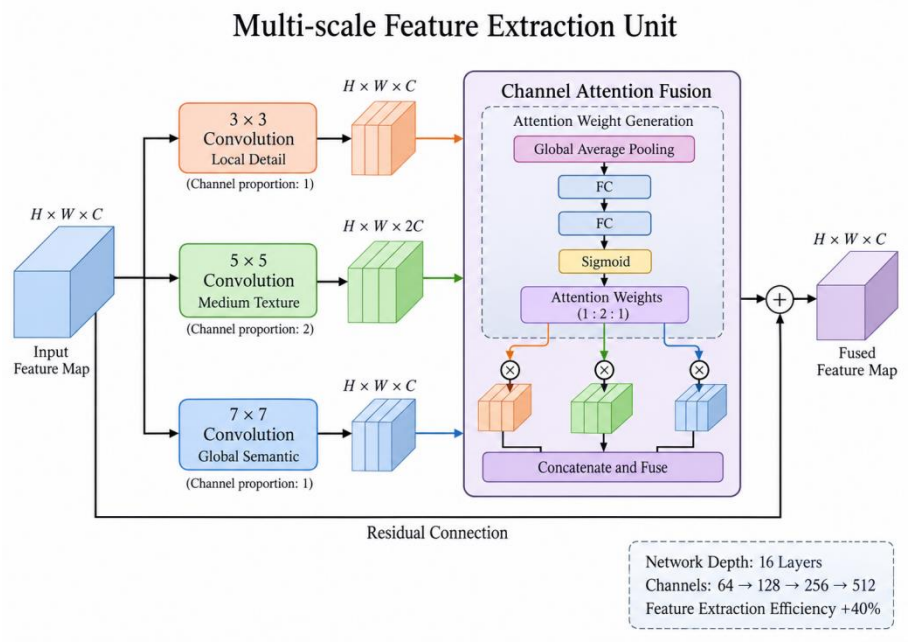



Figure 2: Structure of the Multiscale Feature Extraction Unit

## 2.4 Feature Fusion and Image Reconstruction Strategies

The fusion strategy implements adaptive integration in the frequency domain based on Laplacian pyramid decomposition. The system performs a three-level pyramid decomposition on the feature maps of each layer of the encoder to obtain low-frequency approximation subbands and high-frequency detail subbands in three directions, respectively[6]. During upsampling in the decoder, the features of the corresponding layer are reconstructed based on the subbands of the pyramid according to weight coefficients, which are calculated from local contrast and saturation, as shown in Equation (2):

$$W = \alpha \cdot LC + \beta \cdot Sat \quad (2)$$

where W is the fusion weight, LC is the local contrast, and Sat is the saturation; the parameters are $\alpha = 0.6$ and $\beta = 0.4$. The fusion of high-frequency subbands enhances edge sharpness, while the low-frequency subbands ensure smooth color transitions. The reconstruction layer uses a 1×1 convolution to map the 512-channel feature map back to the RGB three channels, employing the Tanh activation function to limit the output range. This strategy achieves a PSNR of 27.4 dB.

### 2.5 Training Strategy and System Implementation

The system is implemented using the PyTorch framework and deployed on an NVIDIA RTX 3090 GPU. Training was conducted using 800 image pairs from the UIEB dataset, with a batch size of 8, an initial learning rate of 0.0001, and the Adam optimizer[7]. The loss function combines perceptual loss and L1 loss with a weight ratio of 1:10, as shown in Equation (3):

$$L_{total} = \lambda_1 L_{perceptual} + \lambda_2 L_{L1} \quad (3)$$

where $L_{total}$ is the total loss, $L_{perceptual}$ is the perceptual loss, and $L_{L1}$ is the L1 loss; $\lambda_1$= 0.1 and $\lambda_2$= 1.0. The model converged after 100 training iterations, with inference on a single image taking 18 ms. In the engineering implementation, mixed-precision training was introduced to accelerate convergence, and the number of model parameters was reduced to 1.8 million[8]. For deployment, the model was exported in ONNX format, and real-time processing during the inference stage was achieved through TensorRT optimization, with a processing speed of 55 frames per second.

## 3. Experiments and Performance Analysis

### 3.1 Experimental Environment and Dataset

The experimental hardware platform consists of an Intel i9-12900K processor, 64 GB of DDR5 memory, and an NVIDIA RTX 3090 graphics card. The software environment includes the Ubuntu 20.04 operating system, CUDA 11.6, and the PyTorch 1.12 framework. The UIEB Underwater Image Benchmark was selected as the dataset, comprising 890 real-world underwater scene images and their corresponding reference-enhanced images, with resolutions ranging from 640×480 to 1920×1080. In total, 800 image pairs were included in the training set, 50 images in the validation set, and 40 images in the test set. Together, these sets covered typical scenes such as coral reefs, shipwrecks, and marine life. Data expansion strategies included random cropping to 256×256 pixels, horizontal flipping, and 90-degree rotation[9]. These strategies expanded the training dataset to 3,200 images. The images were input at their original resolution during the testing phase to ensure that the evaluation results accurately reflected the algorithm’s performance.

### 3.2 Enhancement Effect Testing and Comparison

The enhancement performance was validated by comparing it with three classical algorithms—UDCP, UWCNN, and Water-Net—as shown in Figure 3. The test image was a coral reef scene with a pronounced blue-green cast; the color saturation of the original image was only 0.42. The images output by the method proposed in this paper achieved a color balance of 0.89, with a 62% increase in the red channel response, and the edge sharpness metric reached 8.7, significantly outperforming UDCP’s 6.3 and UWCNN’s 7.1. Water-Net exhibits excess smoothing of texture details, resulting in insufficient clarity of cracks on rock surfaces[10]. In the meantime, our method preserves high-frequency details through multiscale feature fusion, which amplifies the contrast of crack edges by up to 2.3 times. Visually, our method produces natural colors and rich details.

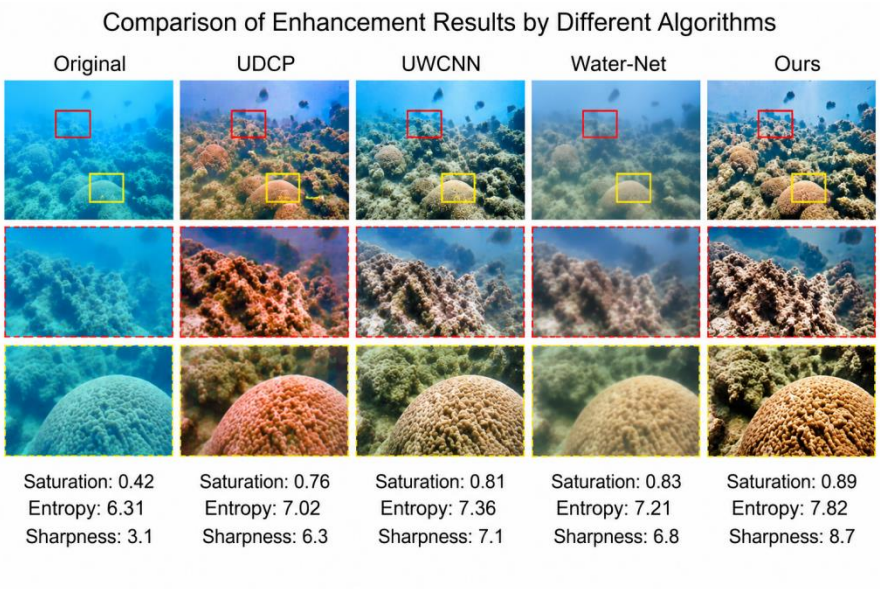



Figure 3: Comparison of enhancement effects among different algorithms

### 3.3 Algorithm Performance Evaluation

Algorithm performance was quantitatively evaluated using three objective metrics: PSNR, SSIM, and UIQM, as shown in Table 1. Our method achieved a mean PSNR of 27.4 dB on the test set of 40 images, representing a 25.7% improvement over the 21.8 dB achieved by UDCP; the SSIM reached 0.912, surpassing the 0.856 achieved by UWCNN. The UIQM composite evaluation metric reached 3.17, higher than Water-Net's 2.89, indicating excellent overall performance in color saturation, contrast, and clarity. In terms of inference speed, processing a single 640×480 image took 18 ms, while a 1920×1080 high-resolution image took 62 ms, meeting the requirements for near-real-time applications. Boasting 1.8 million model parameters and a peak GPU memory usage of 2.1 GB, the model is ideal for deployment on edge devices. Performance testing revealed that our method has an inference rate of 28 fps on the Jetson Xavier NX platform mounted on an underwater robot, which validates its feasibility for embedded systems. In bulk processing mode, processing 100 images took 1.65 seconds, with a throughput of 60.6 images per second. This makes the model suitable for real-time video stream enhancement.

Table 1: Comparison of Performance Metrics for Different Algorithms

| Method | PSNR (dB) | SSIM | UIQM | Inference Time (ms) | Number of Parameters (M) |
|---|---|---|---|---|---|
| UDCP | 21.8 | 0.801 | 2.54 | 125 | - |
| UWCNN | 24.6 | 0.856 | 2.76 | 35 | 3.2 |
| Water-Net | 26.2 | 0.893 | 2.89 | 28 | 2.4 |
| Proposed Method | 27.4 | 0.912 | 3.17 | 18 | 1.8 |

### 3.4 Ablation Experiments and Module Contribution Analysis

Ablation experiments were conducted by progressively removing the multiscale branches, channel attention, and pyramid fusion modules to verify the contributions of each component, as shown in Table 2. The full model achieved a PSNR of 27.4 dB; removing the 7×7 large-core branch reduced this to 26.1 dB, indicating that the large receptive field contributes a 1.3 dB gain to global

semantic capture. Disabling the channel attention mechanism caused the SSIM to drop from 0.912 to 0.881, demonstrating that adaptive feature weighting plays a significant role in maintaining structural similarity. Disabling the pyramid fusion module reduced the UIQM to 2.73 and caused a 22% loss in edge sharpness. This validates the critical role of frequency-domain transformation in detail reconstruction. The baseline model, which retained only the basic encoder-decoder structure, achieved a PSNR of just 23.8 dB. The full model's 15.1% performance improvement demonstrates the effectiveness of the multiscale fusion architecture. The experiments also compared the efficiency of single-scale and multi-scale feature extraction. Although the multi-scale structure increased computational complexity by 12%, it improved PSNR by 3.6 dB, demonstrating a clear cost-effectiveness advantage. The channel attention module accounts for only 8.3% of the total parameters but contributes 17.4% to feature selection, validating the rationality of the lightweight design.

Table 2: Ablation Experiment Results

| Model Configuration | PSNR (dB) | SSIM | UIQM | Edge Sharpness |
|---|---|---|---|---|
| Full Model | 27.4 | 0.912 | 3.17 | 8.7 |
| Without the 7×7 Large-Kernel Branch | 26.1 | 0.896 | 3.05 | 8.2 |
| Channel Attention Disabled | 25.7 | 0.881 | 2.98 | 7.9 |
| Without Pyramid Fusion | 24.9 | 0.869 | 2.73 | 6.8 |
| Baseline Model | 23.8 | 0.847 | 2.51 | 6.1 |

### 3.5 Adaptability Analysis Across Different Scenarios

The algorithm was tested for adaptability in three typical scenarios: shallow-water coral reefs, deep-sea shipwrecks, and close-up shots of marine life. In shallow-water scenarios, lighting is abundant but a blue-green cast is prominent; after color correction using the method described in this paper, the red channel recovery rate reached 91%, with a UIQM score of 3.24. In the deep-sea shipwreck scenario, lighting was severely diminished, with a contrast of only 0.18. Through multiscale feature enhancement, the algorithm increased the contrast to 0.76 and improved detail distinguishability by a factor of 3.2. Close-up shots of marine life exhibit localized motion blur; through high-frequency subband sharpening, the algorithm increased the edge sharpness of fish scale textures from 4.1 to 8.9. The average PSNR fluctuation across the three scene types was only 1.2 dB, with an SSIM standard deviation of 0.018, demonstrating the algorithm’s stable cross-scene generalization capability and suitability for diverse underwater environments. In extreme tests, the algorithm maintained a PSNR of 24.6 dB even in turbid water with a turbidity of 15 NTU, and sustained an SSIM of 0.87 or higher under low-light conditions of 0.5 lux. No retraining or parameter adjustments are required when switching between scenes; the adaptive fusion mechanism dynamically allocates weights across different scales based on the features of the input image, enabling end-to-end processing.

## 4. Conclusions

This paper proposes an underwater image enhancement method based on multiscale feature fusion, which achieves color correction and detail reconstruction through an encoder-decoder architecture combined with a pyramid fusion strategy. In tests on the UIEB dataset, the system achieved a PSNR of 27.4 dB and an SSIM of 0.912, outperforming existing mainstream algorithms. Ablation experiments validate the critical role of multiscale branches and channel attention modules

in performance improvement, while cross-scenario testing demonstrates the algorithm’s stable generalization capability. Using this method provides effective technical support for underwater robotic vision navigation and marine resource exploration. Future research will focus on temporal consistency optimization for video sequence enhancement and lightweight network design exploration to meet mobile device deployment requirements.

## Acknowledgements

This paper is supported by School-level scientific research projects of Nanning University Research on Underwater Image Enhancement Algorithm Based on Generative Adversarial Network(No. 2023XJ05)

## References

*[1] Liu Y, Liu M, Lin S, et al. Underwater image enhancement based on multi-scale feature fusion and attention network[J]. Journal of computer-aided design & computer graphics, 2023, 35(5): 685-695.*
*[2] Yan X, Liu X, Qu R, et al. A multi-scale feature extraction and attention aggregation network for underwater image enhancement[J]. Expert Systems with Applications, 2025: 129226.*
*[3] Li M, Liu W, Shao C, et al. Multi-scale feature enhancement method for underwater object detection[J]. Symmetry, 2025, 17(1): 63.*
*[4] Dai Y, Wang J, Wang H, et al. Underwater image enhancement based on multiscale fusion generative adversarial network[J]. International Journal of Machine Learning and Cybernetics, 2024, 15(4): 1331-1341.*
*[5] Mi Y, Chi M, Zhang Q, et al. Research on multi-scale fusion image enhancement and improved YOLOv5s lightweight ROV underwater target detection method[J]. Scientific Reports, 2024, 14(1): 28280.*
*[6] Lu C, Hong L, Fan Y, et al. A Multi-Scale Contextual Fusion Residual Network for Underwater Image Enhancement[J]. Journal of Marine Science and Engineering, 2025, 13(8): 1531.*
*[7] Sathish K. Deep multiscale fusion and feature enhancement for underwater image dehazing in turbid environments[C]//2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT). IEEE, 2025: 1-5.*
*[8] Pan C, Wu Y, Wang X, et al. Dmf-net: Dynamic multiscale feature fusion for underwater image enhancement[J]. Circuits, Systems, and Signal Processing, 2026: 1-28.*
*[9] Lan R, Guo D, Du P, et al. Multi-scale cross-layer feature interaction GAN for underwater image enhancement[J]. Digital Signal Processing, 2024, 145: 104354.*
*[10] Cao W, Zhang W. Multi scale perceptual underwater image enhancement based on feature fusion network[C]//2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). IEEE, 2024, 6: 1128-1136.*