# Underwater Image Enhancement Using a U-Net Enhanced GAN with Multi-Scale Feature Fusion and Channel-Spatial Attention

Baoyuan Liu, Bin Guo*
School of Electronics and Electrical Engineering, Cangzhou Jiaotong College, Huanghua, 061199, China
E-mail: yongboyb@163.com
*Corresponding author

**Keywords:** Image processing, GAN, Spatial channel attention, multi-scale features, U-Net

**Received:** August 12, 2025

*This study proposes an innovative approach to underwater image processing to counteract the prevalent issue of detail loss associated with current techniques. By enhancing the discriminator within the generative adversarial network through the incorporation of a U-Net, the method aims to preserve the integrity of generated image details. Subsequently, it integrates multi-scale feature fusion and spatial channel attention mechanisms into the collaborative design of the image processing framework to reinforce the expression and reconstruction capabilities of key features. Ultimately, performance validation and analysis were carried out on the Underwater Dark, Underwater ImageNet, and Underwater Scenes datasets, with the primary evaluation metrics including PSNR, SSIM, and Fréchet distance. The experimental results demonstrated that the SSIM of the proposed method across the three datasets was 0.82, 0.85, and 0.83, respectively, which represents an improvement of 0.17, 0.12, and 0.11 over the unimproved methods. Its PSNR maintained the highest values of 28.9, 30.1, and 29.8 across all three datasets, showing an enhancement of 7.9, 7.6, and 7.8 over the baseline models. The Fréchet distance remained the lowest at 22.0, 19.0, and 20.0, indicating a reduction of 21.0, 18.0, and 19.0 compared to the baseline methods. These results indicate that the proposed image processing technique based on U-Net and multi-scale feature fusion effectively solves the problem of detail loss in underwater scenes, has strong generalization ability, and provides feasibility for complex underwater imaging tasks.*

*Povzetek: Študija predstavi izboljšano GAN-metodo za izboljšanje podvodnih slik, ki z U-Net diskriminatorjem ter večnivojskim združevanjem značilk in pozornostnimi mehanizmi bolje ohrani podrobnosti in na več podatkovnih zbirkah doseže višji PSNR/SSIM ter nižjo Fréchet razdaljo kot osnovni modeli.*

## 1 Introduction

Images are one of the media for conveying information. However, traditional Image Processing (IP) techniques exhibit significant limitations in dealing with complex degradation factors in real-world scenarios, such as cross-scale structural distortion and multivariate noise coupling, leading to issues including loss of image details [1-2]. Furthermore, even when employing intelligent methods such as artificial neural networks for multimedia information extraction, challenges regarding robustness persist when confronted with complex degradations like cross-scale structural distortion and coupled multi-noise interferences [3]. Therefore, how to break through the bottleneck of IP technology and develop robust IP methods that can adapt to complex degradation scenarios has become a key issue that urgently needs to be addressed in the current IP field. In recent years, the increasing demand for high-definition imaging and complex scene analysis has driven the evolution of IP technology towards higher accuracy and stronger robustness [4-5]. Therefore, many scholars have conducted research on IP technology. Ilesanmi et al. developed an advanced IP system scheme built on Artificial Intelligence (AI) to address the efficiency and accuracy bottlenecks faced by traditional IP technology in various fields, thereby improving image recognition accuracy [6]. Zhao et al. proposed a new strategy to transform Retinex decomposition into a generative problem to address low contrast and blurry details in Low Light Images (LLIs) that affect human eye recognition. This strategy improved the decoupling effect of dual components in Retinex decomposition and optimized the performance of LLI enhancement [7]. Zhang et al. designed a blind motion deblurring technique based on Deep Learning (DL), which solves the problem of traditional non-blind methods relying on fuzzy kernel estimation and being difficult to cope with complex fuzzy changes [8]. Lei et al. proposed a combination of large-scale pre-training models and language models to address the noise and blur issues in CT images in medical imaging, thereby enhancing the clinical application potential of DL in medical imaging [9]. Zhai et al. put forth a LLI enhancement quality evaluation system built on multidimensional indicators to address the issue of structural damage and noise-induced distortions in LLI enhancement algorithms, as well as the lack of effective quality assessment methods. This study enhanced the

objective evaluation capability of low-light IP quality [10].

In DL-based IP methods, Generative Adversarial Networks (GANs) have shown breakthrough progress in tasks including image super-resolution and denoising through adversarial learning mechanisms [11]. Suganyadevi et al. proposed a multi-disease IP and analysis framework based on DL to address the issue of insufficient accuracy in traditional medical image analysis, thereby improving disease recognition accuracy and achieving intelligent optimization of medical imaging diagnosis [12]. Abdou et al. put forth a medical image comprehensive analysis framework based on Convolutional Neural Networks (CNN) to address the training efficiency, data annotation, and algorithm limitations of DL in medical imaging applications, thereby improving the accuracy of disease diagnosis [13]. Zhou et al. proposed a hybrid contrastive learning regularization network to address the light absorption, scattering, and Color Distortion (CD) caused by complex environments in underwater image enhancement. This improved the model's generalization ability and significantly optimized the visual quality of underwater images [14]. Zhang et al. proposed a GAN-based IP and optimization method to address the issues of realism and completeness in image generation and restoration. By employing collaborative training of generators and discriminators along with the application of global and local patching techniques, they significantly enhanced the authenticity of generated images and efficiently repaired missing regions, thus providing an effective solution for practical applications in the field of IP [15]. However, due to inherent defects such as gradient instability, mode collapse, and receptive field locality, it still faces the challenge of insufficient generation quality in tasks such as image enhancement and restoration.

In summary, existing research in IP has explored various aspects and achieved commendable outcomes. However, there is still a gap in addressing the comprehensive issues of detail loss or CD in underwater IP. Consequently, this study proposes a collaborative processing framework that employs a U-shaped Network (U-Net) to enhance GANs and integrates Multi-Scale Feature Fusion (MSFF) along with attention mechanisms to jointly optimize feature representation, aiming to tackle the problems of detail loss and CD in underwater images. The innovation of the research lies in two aspects. Firstly, the U-Net is utilized to improve GAN and restore image structure and details. Secondly, MSFF is introduced to extract multi-scale features and generate clearer and more realistic images, thereby improving the overall visual quality of underwater images and constructing a new underwater IP technology model. The study aims to introduce a U-Net discriminator, enhance structural consistency constraints, and combine multi-scale features with channel space attention mechanisms to synergistically improve the detail preservation and generalization performance of underwater image enhancement.

Based on the above relevant studies, Table 1 summarizes the research theme, main index methods, and shortcomings of relevant studies.

Based on the above analysis, although current research has achieved good results in general IP tasks such as denoising, deblurring, and weak light enhancement, the performance is not ideal when facing complex underwater degradation factors such as detail loss and CD. At the same time, facing the challenges of underwater imaging scenes such as light absorption, scattering, and noise interference, many existing methods also have bottlenecks in structural consistency, feature representation, generalization ability, and other aspects. Therefore, this study innovatively introduces a U-Net-enhanced GAN integrated with MSFF and channel-spatial attention mechanisms to adaptively restore image details and enhance color fidelity. A novel collaborative processing framework is proposed by combining U-Net, GAN, and MSFF, so that the proposed model can better adapt to the characteristics of underwater environments and achieve efficient and accurate image enhancement.

Table 1: Summary of relevant information of relevant studies

| Author | Method | Dateset | Result | Reported results | Shortcomings |
|---|---|---|---|---|---|
| Ilesanmi et al. [6] | AI-based IP system optimization | / | Improved image recognition accuracy | / | Focuses on general denoising, lacks specialized design for underwater image enhancement. |
| Zhao et al. [7] | Retinex decomposition transformed into a generative problem | Low-light Images | Optimized low-light enhancement performance | PSNR: ~24.5 dB SSIM: ~0.85 | Not designed for underwater color correction and scattering problems. |
| Zhang et al. [8] | DL-based blind motion deblurring | Natural Images | Solved the problem of complex blur variation | PSNR: 28-32 dB | Lacks physical model for underwater degradation; insufficient underwater validation. |
| Lei et al. [9] | Large-scale pre-trained and language models | CT Medical Images | Enhanced clinical application potential | PSNR: >32 dB | Domain-specific (medical), not targeted at underwater optical distortions. |
| Zhai et al. [10] | Multi-dimensional quality assessment | Low-light Images | Improved objective | / | A quality assessment framework, not an |

| | system | | evaluation capability | | underwater enhancement method. |
|---|---|---|---|---|---|
| Suganyadevi et al. [12] | DL-based multi-disease analysis framework | Medical Images | Improved disease recognition accuracy | / | Application-specific (medical), lacks generalization to underwater scenes. |
| Abdou et al. [13] | Comprehensive CNN-based analysis framework | Medical Images | Improved diagnostic accuracy | / | Employs conventional architecture, lacks modern mechanisms like attention. |
| Zhou et al. [14] | Hybrid contrastive learning regularization | Underwater Datasets | Enhanced model generalization and visual quality | PSNR: ~26.2 dB SSIM: ~0.88 | Does not integrate MSFF or a dedicated attention module. |
| Zhang et al. [15] | GAN-based processing and optimization | General Images | Enhanced image realism and completeness | PSNR: ~25.0 dB SSIM: ~0.82 | A general-purpose GAN method, not optimized for complex underwater degradation. |
| This work | U-net+GAN+MSFF with Channel-Spatial Attention | Underwater dark, imagenet, scenes | Effectively improves detail loss and CD in underwater images. | PSNR:~30.1, SSIM: ~0.85 FID:~22.0 | / |

## 2 Methods and materials

This study first introduces the GAN, clarifies its shortcomings, and integrates the U-Net for improvement. Subsequently, a combination of MSFF and improved GAN is introduced, and a GAN-MSFF IP technology framework is designed.

### 2.1 Improved GAN based on U-Net

Traditional IP technology is prone to problems such as loss of details and CD in the face of complex degradation factors in real scenes, which can affect the quality and usability of images. However, the development of DL has brought revolutionary changes to IP. GAN has shown breakthrough progress in IP tasks through its unique adversarial training mechanism. GAN includes a generator and a discriminator, which are trained through adversarial training. This alternating optimization adversarial training mechanism makes it adept at generating complex, high-dimensional data and multimodal samples. However, due to the severe oscillation of the loss function when training, the generated image quality is unstable, and the local perceptual characteristics of its convolution operation make it difficult for the generator to capture global structural information. Therefore, this study utilizes the U-Net to improve GAN. U-Net is a classic full CNN. Its symmetrical U-shaped structure and skip connection design enable it to effectively fuse Feature Maps (Fmap) of the same size, and alleviate the flow of information during gradient updates and forward propagation [16]. Its structure is shown in Fig.1.

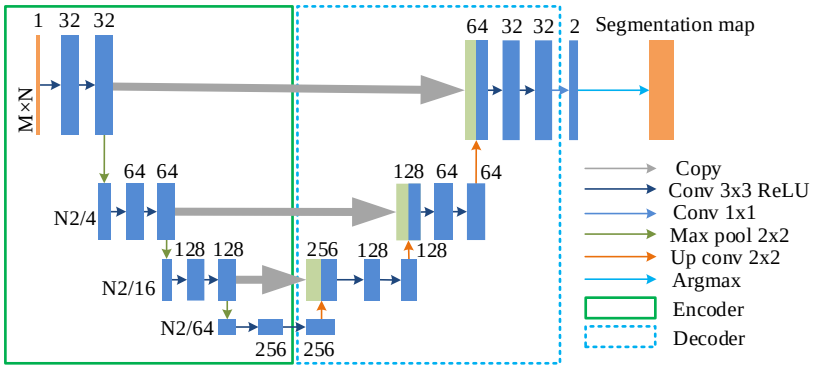



Figure 1: U-Net structure diagram

In Fig.1, the encoder of the U-Net includes multiple convolutional and pooling layers, which are utilized to gradually extract features and reduce resolution. The decoder includes multiple deconvolutions and upsampling layers, taken to gradually restore resolution. The downsampling operation of the pooling layer is shown in equation (1).

$$Y_l = D(Y_{l-1}), l = 1, 2, \ldots, L \quad (1)$$

In equation (1), $D(\cdot)$ is the downsampling operator, $Y_l$ is the downsampled low resolution Fmap, and $l$ is the number of layers. Maximum pooling or average

pooling in the encoder reduces the Fmap size by half, and then restores it through upsampling transpose convolution, as shown in equation (2).

$$\overline{Y_l} = U(Y_{l-1}), l = L, L-1, \ldots, 1 \quad (2)$$

In equation (2), $U(\cdot)$ is the upsampling operator, and $\overline{Y_l}$ is the upsampled high-resolution Fmap. Jumping connection is the process of connecting the Fmaps of the encoder's corresponding layer with those of the decoder, achieving feature fusion between shallow and deep networks without losing edge features, thus preserving more dimensional and positional information. The jumping connection action is shown in equation (3).

$$F_l = Y_l + R(Y_{l+1}), l = L-1, L-2, \ldots, 0 \quad (3)$$

In equation (3), $Y_l$ is the input Fmap of layer $l$, $R(\cdot)$ is the fusion function, and $F_l$ is the fused Fmap. In actual U-Net implementation, the concatenated Fmaps will undergo further convolution processing, which is combined with convolution processing as shown in equation (4).

$$Z_l = C(F_l), l = 0, 1, \ldots, L \quad (4)$$

In equation (4), $C(\cdot)$ is the convolution operation. $Z_l$ is the convolved Fmap, which depends on the quantity of channels in the input and output Fmaps, as well as the size parameters of the convolution kernel. The encoder-decoder structure and skip connections in U-Net can capture contextual and detailed information, while GAN generates realistic images through adversarial training [17-18]. This study adopts the U-Net for the discriminator of GAN, providing more accurate pixel-level feedback to assist in improving the details of the generator [19]. Its structure is shown in Fig.2.

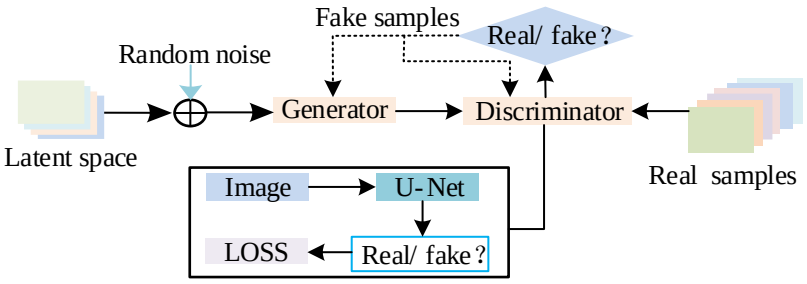



Figure 2: Structure of GAN discriminator using U-Net structure

In Fig.2, the generator receives random noise vectors to generate samples that simulate real data, while the discriminator determines whether the input data are real or generated. The U-Net is utilized to improve the discriminator module in GAN, which includes an encoder and a decoder. The encoder judges the overall authenticity, while the decoder compares pixel by pixel to calculate the reconstruction error. This design can preserve local detail information while capturing global true and false features.

## 2.2 IP technology based on improved GAN combined with MSFF

An improved GAN based on U-Net can capture the overall structural semantic information of images, thereby improving the problem of local detail loss in generated images. However, it has shortcomings in extracting and fusing features of different scales, and has poor segmentation performance in image quality and spatial consistency. Therefore, this study further designs an IP technology for GAN-MSFF to enhance the extraction and feature fusion at different scales, maintain overall semantic consistency in the generated images, and generate clearer and more realistic images. MSFF can extract features at different levels and resolutions of the model and integrate them through specific algorithms to achieve a comprehensive representation from local details to global semantics [20-21]. The extraction of MSFF is exhibited in Fig.3.

In Fig.3, MSF extraction requires preprocessing of the image before performing Multi-Scale Decomposition (MSD), followed by feature extraction and fusion, and finally reconstructing the result and outputting the Fmap. In the MSD, the low-frequency layer obtains the global brightness trend through Gaussian pyramid decomposition. The high-frequency layer uses a wavelet transform to extract multi-band information, followed by Laplace pyramid decomposition, as shown in equation (5).

$$E_l = Y_l - U(D(Y_l)), l = 1, 2, \ldots, L \quad (5)$$

In equation (5), $E_l$ is the residual detail map of layer $l$. The image undergoes MSD for feature extraction and fusion, and in the feature fusion stage, the collaborative optimization of MSFs is achieved through weighted averaging, attention mechanism, and cross layer connections. The weighted average is shown in equation (6).

$$Y_{final} = \sum_{l=0}^{L} \alpha_l \overline{Y_l} \quad (6)$$

In equation (6), $\alpha_l$ is the interlayer weight. $Y_{final}$ is the final reconstructed high-resolution image. The weighted average mechanism achieves preliminary alignment at the feature level by dynamically allocating weights. It makes the model adaptively adjust the contribution ratio of different source features based on their actual importance, effectively solving the problem of traditional average pooling ignoring the importance differences between features due to a "one size fits all

approaches". Subsequently, the features are embedded into the Channel Space Dual Attention (CSDA) module for multi-level feature filtering. The structure of CSDA is shown in Fig.4.

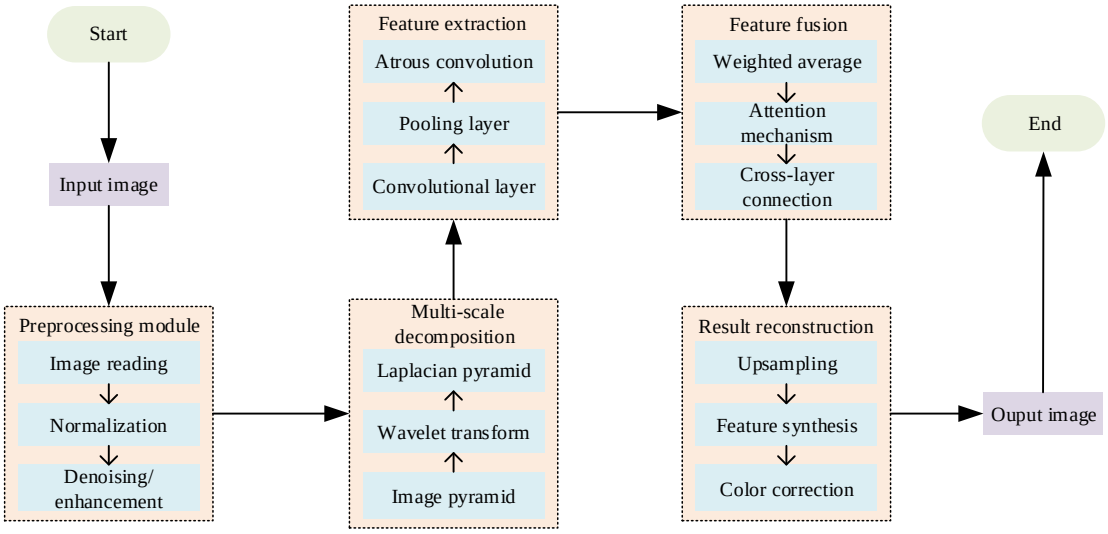



Figure 3: Flowchart of MSF extraction

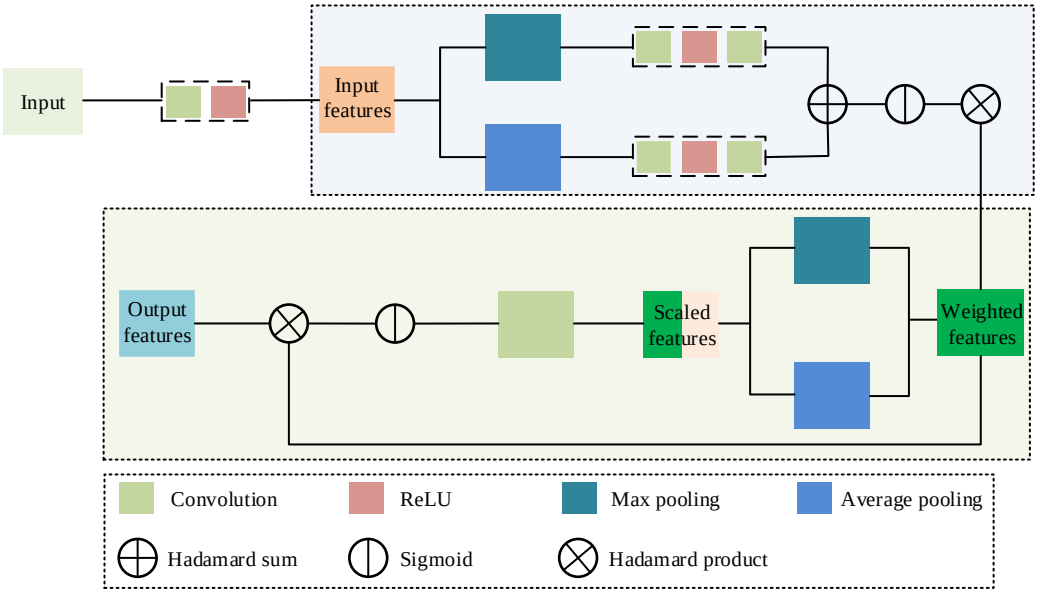



Figure 4: CSDA module structure

In Fig.4, the input features undergo convolution and other operations before generating weights through Sigmoid's activation function, which is used to weight the original features to highlight important channel information. The spatial attention part performs attention calculation on the Fmap in the spatial dimension and learns the importance of different positions in space. Finally, corresponding weight adjustments are made to enhance or suppress features at different spatial positions, as shown in equation (7).

$$\begin{cases} A_s = \mathrm{Softmax}(W_s \cdot \mathrm{GlobalAvgPool}(Z_l)) \\ A_c = \mathrm{Sigmoid}(W_c \cdot Z_l) \\ Z_l^{'} = Z_l \otimes A_s \otimes A_c \end{cases} \tag{7}$$

In equation (7), $W_s$ and $W_c$ are learnable weights. GlobalAvgPool is global average pooling. $\otimes$ is element level multiplication. $A_s$ and $A_s$ are spatial and channel attention maps. $Z_l^{'}$ is the Fmap weighted by attention. The selected features need to be bridged between shallow details and deep semantics using residual structures, as shown in equation (8).

$$\bar{\mathrm{I}} = \sum_{p\in\Omega} \mathrm{W}_p \cdot \mathrm{Y}_{final}(p) \quad (8)$$

In equation (8), $\mathrm{W}_p$ is the weight matrix of the $p$-th image block, and $\bar{\mathrm{I}}$ is the reconstructed image block. After completing the feature fusion of the image, the feature results are reconstructed and the Fmap is output. In the result reconstruction section, the image needs to complete color space conversion and calculate the total variational loss, as shown in equation (9).

$$\begin{cases} \mathrm{I}_{final} = \mathrm{T}(\bar{I}) \\ \mathrm{K}_{TV}(\mathrm{I}_{final}) = \sum_{ij}((\mathrm{I}_{final}(i+1,j) - \mathrm{I}_{final}(i,j)^2 + (\mathrm{I}_{final}(i,j+1) - \mathrm{I}_{final}(i,j))^2) \end{cases} \quad (9)$$

In equation (9), $\mathrm{T}(\cdot)$ is the color space conversion function. $\mathrm{I}_{final}$ is the final output color high-resolution image. $\mathrm{K}_{TV}(\mathrm{I}_{final})$ is the total variation loss. $\mathrm{I}_{final}(i,j)$ is the value at pixel position $(i,j)$. $\sum_{i,j}$ is the sum of all pixel positions in the image. MSFs not only enhance the discriminator's ability but also assist the generator in generating higher quality images. The fusion strategy combined with improved GAN is shown in Fig.5.

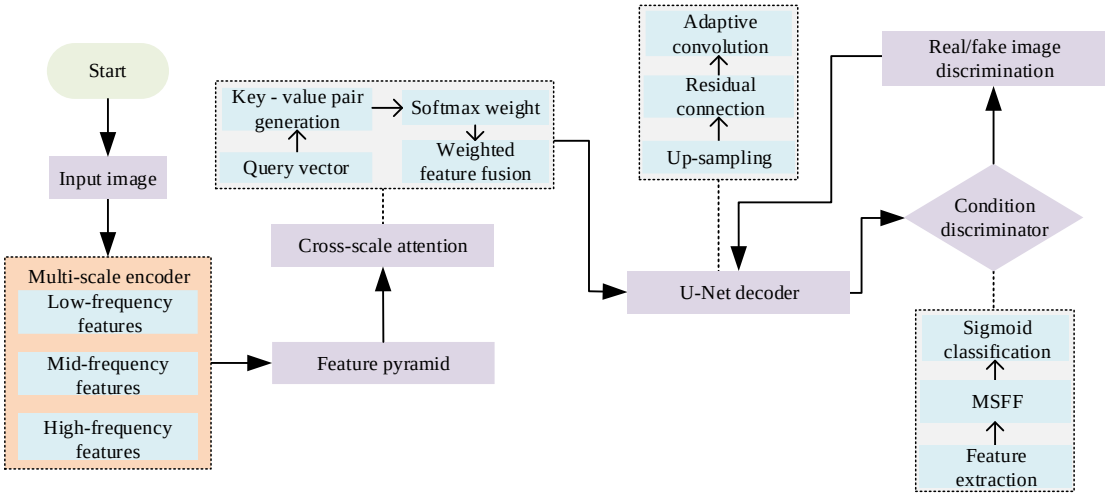



Figure 5: U-net+GAN+MSFF structure diagram

In Fig.5, firstly, the original image is input into a multi-scale encoder, which gradually reduces the image size through a downsampling operation, and then extracts the corresponding feature information of the image at different scales. Secondly, feature pyramid fusion is implemented for features of different scales to integrate multi-level image information, and the obtained features are transmitted to the cross-scale attention module to obtain key information. The processed features are transmitted to the U-Net decoder. It restores the image resolution through upsampling and outputs the generated result of the generator, which is then input to the discriminator along with the original image. The discriminator first extracts features from the input image, and then performs classification operations using MSFF and Sigmoid activation functions to determine whether the input image is a real or generated image. The discrimination result forms a feedback signal that propagates back to the generator. Its function is to guide the parameter update and optimization of the generator, and enhance the quality and authenticity of the generated images. The expression of the conditional adversarial loss function is given by equation (10).

$$K = \mathrm{E}_{\mathrm{Y},\overline{\mathrm{Y}}}\left[\log D(\mathrm{Y},\overline{\mathrm{Y}})\right] + \mathrm{E}_{\mathrm{Y},\overline{\mathrm{Y}}}\left[\log(1 - D(\mathrm{Y},\underline{\overline{\mathrm{Y}}}))\right] \quad (10)$$

In equation (10), $\mathrm{Y}$ is a real high-resolution image, $\overline{\mathrm{Y}}$ is a model generated image, $\underline{\overline{\mathrm{Y}}}$ is a fake image, and $D(\cdot)$ is the discriminator function. The conditional adversarial loss function maximizes the misclassification probability of the discriminator, forcing the generated image to approximate the true distribution at the content level. The complete loss function combination not only includes the basic adversarial loss but also introduces additional perceptual loss and structural similarity loss to address the limitations of pure adversarial training. Its expression is shown in equation (11).

$$\begin{cases} K_{SSIM} = 1 - \mathrm{SSIM}(\mathrm{Y},\overline{\mathrm{Y}}) \\ K_{Perceptual} = \left\| \phi(\mathrm{Y}) - \phi(\overline{\mathrm{Y}}) \right\|_2^2 \end{cases} \quad (11)$$

In equation (11), $\mathrm{SSIM}(\cdot)$ is the Structural Similarity Index (SSIM), $\phi(\cdot)$ is the feature extractor of the pre-trained VGG-19 network, and $\|\cdot\|_2^2$ is the $L_2$-norm. Perceived loss can provide local pixel-level and structural-level constraints, ensuring high standards of detail and overall quality in the generated image. SSIM loss provides a global semantic-level supervised signal, guiding the generator to learn the essential features of data distribution. Therefore, this mixed loss design can not only avoid the problem of pattern collapse in pure adversarial training but also overcome the

blurring artifacts that may occur due to relying solely on pixel-level losses. Combined with the above, the pseudo-code of the proposed method is shown in Table 2.

Table 2: Core implementation of the U-net+GAN+MSFF algorithm

| The pseudo-code of the U-net+GAN+MSFF algorithm |
| --- |
| input torch<br>input torch.nn as nn<br>input torch.nn.functional as F<br>class UnderwaterEnhancementGAN:<br>"""Core implementation of U-net+GAN+MSFF for underwater image enhancement"""<br>def __init__(self, device='cuda'):<br>self.generator = self._build_unet_generator()<br>self.discriminator = self._build_discriminator()<br>self.device = device<br>def _build_unet_generator(self):<br>"""U-Net generator with skip connections for detail preservation"""<br>return UNetWithSkipConnections()<br>def _build_discriminator(self):<br>"""Discriminator with multi-scale feature fusion"""<br>return MultiScaleDiscriminator()<br>def train(self, dataloader, epochs=200):<br>"""Training procedure with adversarial learning"""<br>for epoch in range(epochs):<br>for batch_idx, (real_imgs, target_imgs) in enumerate(dataloader):<br># Generate enhanced images<br>enhanced_imgs = self.generator(real_imgs)<br># Adversarial training<br>d_loss = self._update_discriminator(real_imgs, enhanced_imgs, target_imgs)<br>g_loss = self._update_generator(real_imgs, enhanced_imgs, target_imgs)<br>if epoch % 20 == 0:<br>print(f"Epoch {epoch}: G_loss={g_loss:.4f}, D_loss={d_loss:.4f}")<br>def enhance_image(self, underwater_img):<br>"""Enhance underwater image using trained model"""<br>return self.generator(underwater_img)<br># Usage exampleif __name__ == "__main__":<br>model = UnderwaterEnhancementGAN()<br># Load pre-trained weights<br># model.generator.load_state_dict(torch.load('weights/u_net_gan_msff.pth'))<br>print("U-net+GAN+MSFF model ready for underwater image enhancement!") |

# 3 Results and discussion

This study first introduced the relevant parameters of the experimental setup and analyzed the IP capability of models with different combination structures in ablation experiments. Secondly, the performance before and after improvement was verified, and finally, other IP technologies were introduced for performance comparison.

## 3.1 Experimental setup

To verify the effectiveness of the improved GAN combined with MSFF, this study uses the EUVP dataset proposed by Jahidul Islam et al. and transformed unpaired images into paired datasets using CycleGAN. Finally, performance validation is conducted on the Underwater Dark (UWD), Underwater ImageNet (UWI), and Underwater Scenes (UWS) datasets [22]. Before training, the study implements uniform data preprocessing on the input images. Initially, the image resolution is adjusted to 256×256 pixels and subjected to normalization, linearly scaling the pixel values from the range [0, 255] to [-1, 1]. Concurrently, random horizontal flipping is employed as a data augmentation technique to enhance the model's generalization capability. The EUVP dataset comprises a total of 11,435 image samples, which are randomly divided into training and testing sets in an 8:2 ratio, while the UWD, UWI, and UWS datasets contain 5,550, 3,700, and 2,185 image samples respectively, all of which are utilized for testing purposes. Meanwhile, the validation results are evaluated using SSIM, Peak Signal-to-Noise Ratio (PSNR), and Fréchet Induction Distance (FID) as metrics. The calculation of the FID is designed to assess the similarity between the feature distributions of generated images and real images. To ensure consistency, all input images are resized to a resolution of 299×299 pixels prior to FID computation, and feature extraction is performed using a pre-trained Inception v3 model within the PyTorch environment.

Additionally, the total loss function is a weighted sum of adversarial loss, structural similarity loss, and perceptual loss, with the study configuring the weights to 1.0, 1.0, and 0.1, respectively. This ratio aims to balance the contributions of different losses, following the configuration in reference [14]. The entire experiment is conducted on the Win 11 system, with an Intel Core i7-10875H 2.3GHz processor and software frameworks written in PyTorch 2.1 and Python 3.9. The experimental environment and parameter settings are shown in Table 3.

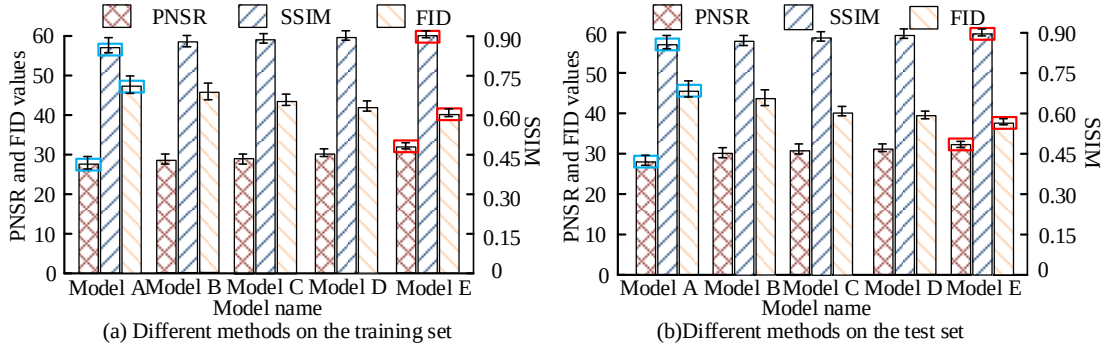
Table 3: Parameter settings

| Parameter name | Parameter setting |
|---|---|
| Learning rate | 0.0002 |
| Batch size | 16 |
| Epochs | 200 |
| Optimizer | Adam |
| GPU | RTX 3090 (24GB) |
| Learning rate scheduler | Cosine annealing |
| Memory usage | 16 GB |

The parameters in Table 3 are obtained through model training and performance control on preprocessed images from three datasets.

## 3.2 Performance verification of U-net+GAN+MSFF model

To validate the effectiveness of the U-net+GAN+MSFF model, this study gradually added modules to measure the independent contribution of each module and analyzed the effects of different modules on IP. Meanwhile, the EUVP dataset was segmented into a training set that accounted for 80% and a testing set that accounted for 20%. The reason was that EUVP, as a comprehensive underwater image dataset, preserved the natural distribution characteristics of the original underwater images. Therefore, based on the original distribution of data, this study objectively evaluated the basic performance of the model without additional data conversion processing. Subsequent experiments would select datasets that had undergone data conversion processing, and through this experimental setup on different types of datasets, comprehensively and accurately demonstrate the effectiveness and superiority of U-net+GAN+MSFF IP technology. The setup of the ablation experimental model includes: GAN, MSFF, U-net+GAN, GAN+MSFF, and U-net+GAN+MSFF, denoted as Models A-E. The validation results of the five models in the two sets are displayed in Fig.6.




(a) Different methods on the training set (b)Different methods on the test set

Figure 6: Test results of various methods in EUVP dataset

In Fig.6 (a), in the training set, the PSNR and SSIM of all models are steadily increasing, indicating that the model's capacity to fit the training data is constantly improving, but the FID fluctuates greatly. Model B has improved PSNR, SSIM, and FID by 0.5, 0.02, and -1.9 compared to Model A, but lower than Model E by 3, 0.07, and 4.8, respectively. In Fig.6 (b), in the test set, the PSNR and SSIM of all models are steadily increasing, with Model B's PSNR increasing by 0.6 compared to Model A, indicating that MSFF can effectively improve high-frequency details. Based on Figs.6 (a) and (b), the red boxes indicate the best results, while the blue boxes show the worst, demonstrating the effectiveness of the model improvements. Model C shows a substantial FID reduction in the test set, signifying a marked enhancement in structural consistency by the U-Net discriminator. Model E outperforms in both training and test sets, indicating that feature selection has effectively mitigated redundant noise. Table 4 details the inference time and resource consumption of different models on the EUVP dataset.

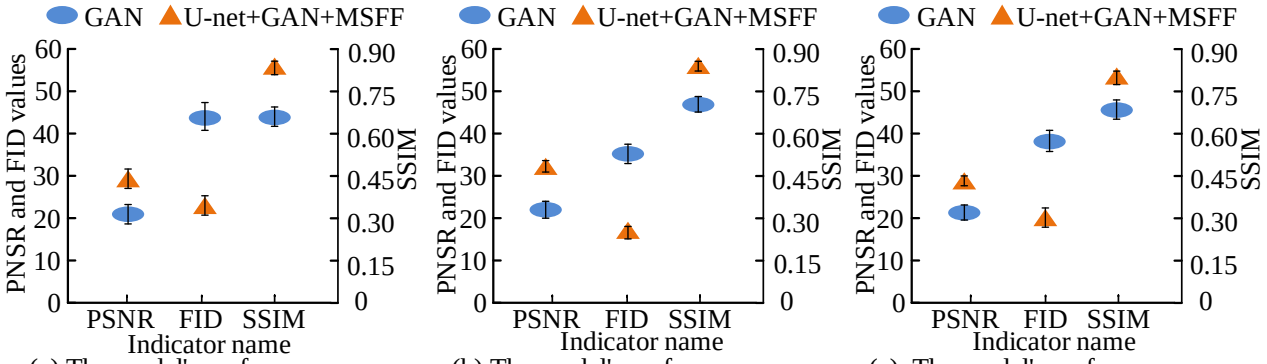
Table 4: Comparison of the inference efficiency and resource consumption of various models on the EUVP dataset

| Model name | Model A | Model B | Model C | Model D | Model E |
|---|---|---|---|---|---|
| Inference time (ms) | 45 | 50 | 54 | 61 | 68 |
| GPU memory usage (GB) | 4.2 | 4.4 | 4.9 | 5.3 | 5.9 |

In Table 4, Model A serves as the baseline with the shortest inference time and lowest resource consumption. Model E exhibits only a 23 ms increase in inference time and 1.7 GB increase in memory usage compared to Model A, yet it significantly outperforms the other four models in terms of PSNR, SSIM, and FID on the EUVP dataset. Therefore, considering comprehensive performance, Model E enhances image enhancement while maintaining high computational efficiency, making it feasible for practical deployment.

To verify the model’s performance, further analysis

is conducted on the IP capabilities before and after improvement in different datasets, including UWD, UWI, and UWSs. The data processing results are shown in Fig.7.



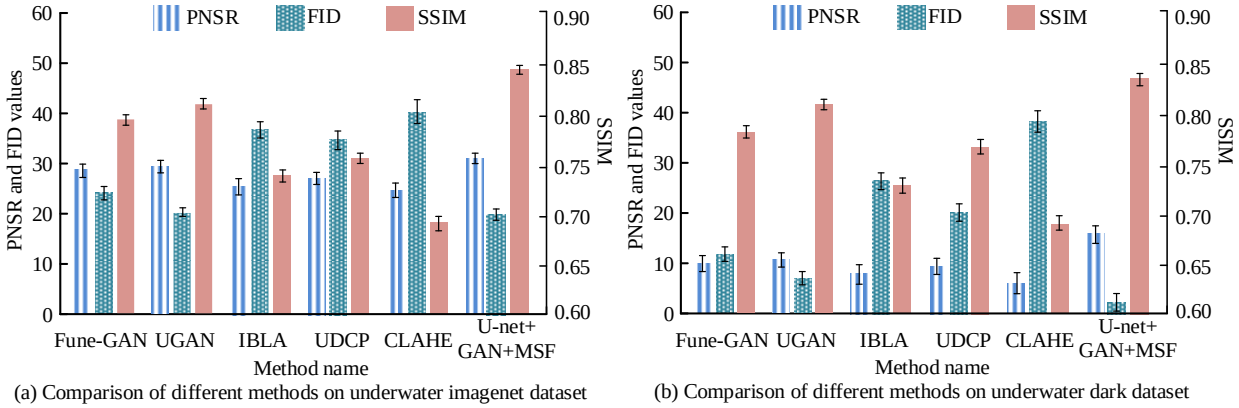

(a) The model's performance on the underwater dark dataset (b) The model's performance on the underwater imagenet dataset (c) The model's performance on the underwater scenes dataset

Figure 7: Test results in three datasets before and after model improvement

In Fig.7 (a), the PSNR, SSIM, and FID of GAN are 22.1, 0.68, and 45, which are 6.8 lower than those of U-net GAN+MSFF, 0.14 lower than SSIM, and 23 higher than FID. In Fig.7 (b), in UWI, the PSNR and SSIM of U-net-GAN+MSFF increase by 5.8 and 0.13, but the FID decreases by 19. In Fig.7 (c), GAN performs poorly, possibly due to the improved model capturing local details through shallow networks and extracting global layout through deep networks in complex scenes, thus avoiding local overfitting. Based on Figs.7 (a)~(c), U-net GAN+MSFF maintains the best performance in all three extreme scenarios, reflecting that the model has a higher priority for optimizing noise sensitive scenarios.

## 3.3 Performance evaluation and comparison of different IP technologies

To further substantiate the efficacy of the U-Net+GAN+MSFF technology, this study incorporates a variety of IP methodologies and validates them across the UWD, UWI, and UWS datasets. Funie-GAN represents an advanced GAN-based model focusing on specific image generation or style transformation tasks. Unsupervised GAN (UGAN) is a variant of GAN capable of training without paired data, learning mappings from the source domain to the target domain without corresponding relationships. Image-based Lighting Adjustment (IBLA) is utilized for enhancing image quality. Unsupervised Domain-Consistent Projection (UDCP) is employed for feature transformation between different domains without the necessity for labeled data. Contrast Limited Adaptive Histogram Equalization (CLAHE) is a widely used image enhancement technique aimed at improving local contrast while curbing noise amplification. The verification outcomes of these diverse IP techniques on the UWD and UWI datasets are depicted in Figure 8.




(a) Comparison of different methods on underwater imagenet dataset (b) Comparison of different methods on underwater dark dataset

Figure 8: Comparison of different IP techniques in UWI and UWD

In Fig.8 (a), the traditional methods of CLAHE and IBLA perform poorly in PSNR, at 23.9 and 24.8. In SSIM, CLAHE performs the worst (0.68), while U-net+GAN+MSFF performs the best (0.82), and it still remains the best in FID. In Fig.8 (b), CLAHE has the worst PSNR performance at 24.5, followed by IBLA at 25.6. Meanwhile, UDCP performs better than CLAHE and IBLA methods, but it is inferior compared to Fune GAN and UGAN. Combining Figs.8 (a) and (b), the traditional method performs the worst in both scenarios,

which may be due to the fact that this method exacerbates local contrast enhancement and noise amplification due to linear processing and non-adaptive enhancement strategies, causing blurry and color distorted generated images. U-net+GAN+MSFF performs the best in both scenarios, with the highest PSNR and SSIM of 30.1 and 0.85 in the standard scenario, and a FID close to the ideal value of 19. To further validate the effectiveness of the algorithm and address the limitations of existing evaluation systems, this study introduces Root Mean Square Error (RMSE), Variance Inflation Factor (VIF), and Entropy as supplementary evaluation indicators. The supplementary indicators of different IP technologies in UWD and UWI are shown in Table 5.

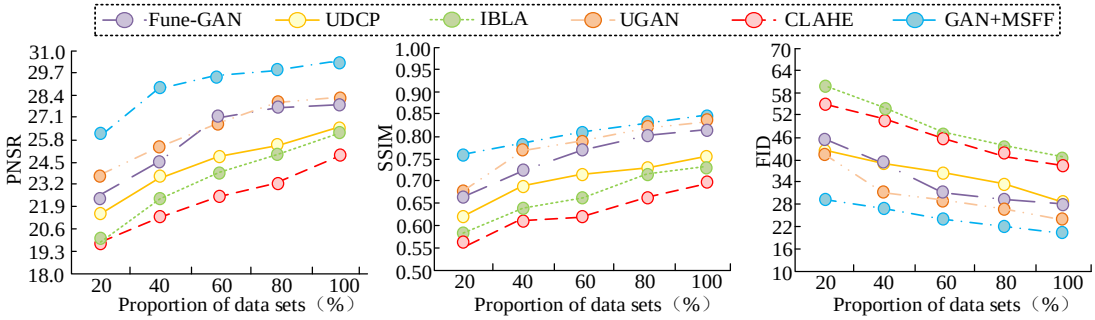
Table 5: Complementary metrics validation of IP techniques on UWD and UWI

| Dataset name | Method name | RMSE | VIF | Entropy |
|---|---|---|---|---|
| UWD dataset | U-net+GAN+MSFF | 18.2 | 0.68 | 6.2 |
| | Fune-GAN | 20.7 | 0.62 | 6.0 |
| | UGAN | 23.1 | 0.55 | 5.8 |
| | IBLA | 25.4 | 0.49 | 6.5 |
| | UDCP | 27.8 | 0.42 | 5.5 |
| | CLAHE | 30.1 | 0.38 | 7.1 |
| UWI dataset | U-net+GAN+MSFF | 15.3 | 0.75 | 6.0 |
| | Fune-GAN | 16.9 | 0.71 | 5.7 |
| | UGAN | 18.7 | 0.63 | 6.2 |
| | IBLA | 20.1 | 0.68 | 5.9 |
| | UDCP | 22.4 | 0.56 | 5.4 |
| | CLAHE | 24.6 | 0.51 | 6.8 |

In Table 5, U-net+GAN+MSFF performs the best in low light scenes, with an RMSE increase of -11.9 compared to CLAHE, indicating small pixel-level differences. Compared with UDCP, the VIF of U-net+GAN+MSFF increases by 56.5%. This indicates that U-net+GAN+MSFF has good structural and texture preservation ability in IP. In standard scenarios, although Fune-GAN performs better than other traditional methods, it is inferior compared to U-net+GAN+MSFF. Meanwhile, UDCP performs the worst in low-light scenes, with the highest RMSE of 30.1 and the lowest VIF of 0.38. This may be because UDCP is not optimized for extreme dark environments. CLAHE exhibits high Entropy of 7.1 and 6.8 in both lighting scenarios, which may be due to its tendency to introduce over-enhancement in diverse data. The results of different IP technologies in UWSs are shown in Fig.9.

In Fig.9 (a), in complex scenarios, the PNSR values of all methods increase with the increase of dataset proportion, while U-net+GAN+MSFF leads other methods at all data proportions. When the dataset proportion is 100%, its highest value is 29.8, which is 1.6 and 2.3 higher than the better performing UGAN and Fune GAN. In Fig.9 (b), the growth trend of U-net+GAN+MSFF is relatively slow, with a numerical difference of only 0.08, which is much lower than the three methods of IBLA, UDCP, and CLAHE with numerical differences of 0.14, 0.12, and 0.13. This may be because traditional methods have failed to generate in complex scenarios. In Fig.9 (c), the change trends of IBLA and UGAN are the most obvious, with differences of 22.0 and 14.7, respectively. When the dataset ratio is 100%, the differences between IBLA, UGAN, and MSFF are 18.0 and 4.7. Combining Figs.9 (a) to (c), the PNSR and SSIM of U-net+GAN+MSFF are the highest at all data ratios, while the FID is the lowest, verifying its strong robustness in small samples. The difference in PNSR and FID changes of GAN-based methods is lower than that of traditional methods, indicating that traditional methods are more sensitive to changes in data volume and have instability in complex scenarios. The supplementary indicators of different IP technologies in the UWS dataset are shown in Fig.10.

In Fig.10 (a), the UDCP method performs the worst in UWSs, with RMSE and Entropy of 25.7 and 5.6, showing an increase of -8.3 and 0.40 in RMSE and Entropy compared to U-net+GAN+MSFF. In Fig.10 (b), U-net+GAN+MSFF performs the best on VIF at 0.72, while CLAHE performs the worst at only 0.47. In Figs.10 (a) and (b), under dynamic scenarios, the performance of U-net+GAN+MSFF still maintains a leading position, while UDCP exhibits a severe failure state. This may be due to the lack of a motion compensation mechanism in UDCP. CLAHE exhibits high Entropy, possibly due to its excessive enhancement of rapidly changing regions.

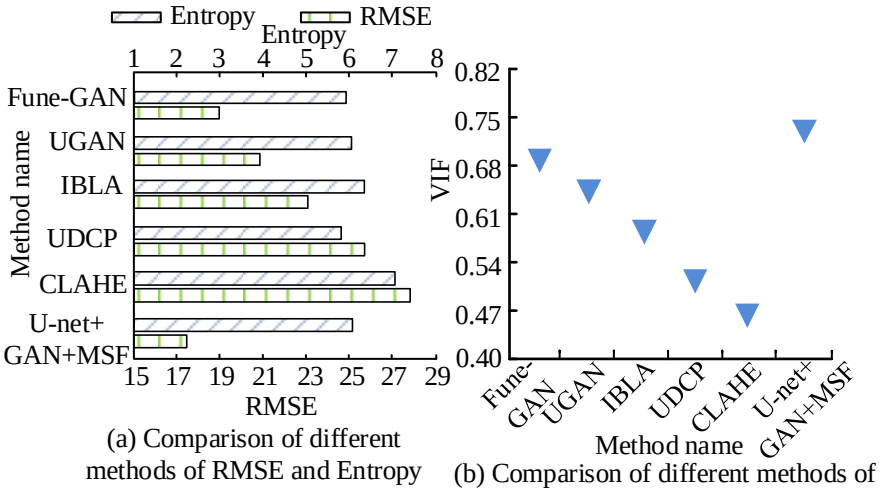



(a) PNSR of different methods in underwater scenes dataset

(b) SSIM of different methods in the underwater scenes dataset

(c) FID of different methods in the underwater scenes dataset

Figure 9: Performance comparison of methods on different scale UWS datasets




(a) Comparison of different methods of RMSE and Entropy on underwater scenes dataset

(b) Comparison of different methods of VIF on underwater scenes dataset

Figure 10: Complementary metrics results for different IP techniques on the UWS

# 4 Discussion

Addressing the issues of detail loss and CD in underwater images, this study proposes an underwater IP technique based on U-net+GAN+MSFF and validates it through experiments on relevant datasets. Experimental results demonstrate that the proposed U-net+GAN+MSFF model significantly outperforms existing IP methods across multiple underwater datasets. This performance advantage stems from its synergistically innovative architectural design: employing U-Net as the discriminator ensures spatial structural consistency and detail integrity in the generated images through skip connections, thereby improving the SSIM metric. The MSFF mechanism effectively integrates features from global semantics to local edges, resulting in exceptional pixel-level reconstruction accuracy as reflected by the PSNR. Simultaneously, the channel-spatial attention module enhances critical features and suppresses noise through dynamic weighting, making the feature distribution of the generated images closer to real data, which is concretely manifested by a significant reduction in the FID value. Most importantly, the model exhibits strong robustness in complex underwater scenarios with low illumination and high noise, benefiting from the targeted optimization of the aforementioned mechanisms against CD and detail loss. However, as the model has only been validated in scenarios with partial motion blur, its performance in extreme conditions requires further investigation. Moreover, the current computational complexity poses a challenge for deployment on edge devices, indicating that future research should focus on lightweight architectures and cross-modal generalization capabilities.

# 5 Conclusion

This study innovatively proposed an IP technology of U-net+GAN+MSFF to address issues such as loss of details or CD in images. It was systematically validated on the UWD, UWI, and UWS datasets. In the ablation experiments on the EUVP dataset, GAN+MFSS maintained high PSNR, SSIM, and low FID in both the training and testing sets. In UWD, the PSNR, SSIM, and FID of U-net+GAN+MFSS were 28.9, 0.82, and 22, which were superior to Fune-GAN and UGAN methods. Its PSNR increased by 2.7 and 1.8 compared to Fune-GAN and UGAN, while FID decreased by 6.0 and 2.0, indicating that the generated image distribution in low light environments was closest to real data. In UWI, the PSNR, SSIM, and FID of U-net+GAN+MFSS were 30.1, 0.85, and 19.0, which were significantly better than traditional methods and GAN-based methods. In UWSs, the PSNR, SSIM, and FID of U-net+GAN+MFSS were 29.8, 0.83, and 20.0 when the dataset ratio was 100%. Their FID decreased by 18.0 and 22.0 compared to traditional methods. This indicates the adaptive advantage of the DL architecture under complex underwater conditions. Research has shown that U-Net+GAN+MSFF has significant advantages in underwater IP and can effectively address issues such as detail loss or CD that occur in traditional methods. However, this study currently focuses on the details of underwater IP or CD, and has not yet deployed and verified the network architecture of edge computing devices. Therefore, future work will focus on developing efficient and lightweight network architectures to satisfy the practical deployment needs of edge devices and other scenarios.

# Fundings

The research is supported by Self funded project of Cangzhou Science and Technology Plan for 2023-2024 (23244101004) “Research on Recognition Algorithm of High-speed Railway Maintenance Tools Based on Convolutional Neural Network”. Funded by Science Research Project of Hebei Education Department (ZC2024119).

# References


[1] Pal S, Roy A, Shivakumara P, Pal U. Adapting a Swin Transformer for License Plate Number and Text Detection in Drone Images. Artificial Intelligence and Applications, 2023, 1(3): 145-154. https://doi.org/10.47852/bonviewAIA3202549

[2] Yang X, Yu Z, Xu L, Hu J, Wu L, Yang C, Zhang Y. Underwater ghost imaging based on generative adversarial networks with high imaging quality. Optics Express, 2021, 29(18): 28388-28405. https://doi.org/10.1364/OE.435276

[3] Liu L, Yang Y. A Study on the Application of New Feature Techniques for Multimedia Analysis in Artificial Neural Networks by Fusing Image Processing. Informatica, 2024, 48(11). https://doi.org/10.31449/inf.v48i11.5851

[4] Ashtiani F, Geers AJ, Aflatouni F. An on-chip photonic deep neural network for image classification. Nature, 2022, 606(7914): 501-506. https://doi.org/10.1038/s41586-022-04714-0

[5] Li J, Xi C, Ke Y. Robustness Prediction of Complex Networks Based on CNN Improved by Graph Representation Learning Operators. Informatica, 2025, 49(21). https://doi.org/10.31449/inf.v49i21.7316

[6] Ilesanmi A E, Ilesanmi T O. Methods for image denoising using convolutional neural network: a review. Complex and Intelligent Systems, 2021, 7(5): 2179-2198. https://doi.org/10.1007/s40747-021-00428-4

[7] Zhao Z, Xiong B, Wang L, Ou Q, Yu L, Kuang F. RetinexDIP: A unified deep framework for low-light image enhancement. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3): 1076-1088. https://doi.org/10.1109/TCSVT.2021.3073371

[8] Zhang K, Ren W, Luo W, Lai W S, Stenger B, Yang M H, Li H. Deep image deblurring: A survey. International Journal of Computer Vision, 2022, 130(9): 2103-2130. https://doi.org/10.1007/s11263-022-01633-5

[9] Lei Y, Niu C, Zhang J, Wang G, Shan H. CT image denoising and deblurring with deep learning: current status and perspectives. IEEE Transactions on Radiation and Plasma Medical Sciences, 2023, 8(2): 153-172. https://doi.org/10.1109/TRPMS.2023.3341903

[10] Zhai G, Sun W, Min X. Perceptual quality assessment of low-light image enhancement. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2021, 17(4): 1-24. https://doi.org/10.1145/3457905

[11] Ross N S, Shibi C S, Mustafa S M, Gupta M K, Korkmaz M E, Sharma V S, Li Z. Measuring surface characteristics in sustainable machining of titanium alloys using deep learning-based image processing. IEEE Sensors Journal, 2023, 23(12): 13629-13639. https://doi.org/10.1109/JSEN.2023.3269529

[12] Suganyadevi S, Seethalakshmi V, Balasamy K. A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval, 2022, 11(1): 19-38. https://doi.org/10.1007/s13735-021-00218-1

[13] Abdou M A. Literature review: Efficient deep neural networks techniques for medical image analysis. Neural Computing and Applications, 2022, 34(8): 5791-5812. https://doi.org/10.1007/s00521-022-06960-9

[14] Zhou J, Sun J, Li C, Jiang Q, Zhou M, Lam K M, Fu X. HCLR-Net: hybrid contrastive learning regularization with locally randomized perturbation for underwater image enhancement. International Journal of Computer Vision, 2024, 132(10): 4132-4156. https://doi.org/10.1007/s11263-024-01987-y

[15] Zhang Y, Xie H, Zhuang S, Zhan X. Image Processing and Optimization Using Deep Learning-Based Generative Adversarial Networks (GANs). Journal of Artificial Intelligence General Science (JAIGS), 2024, 5(1): 50-62. https://doi.org/10.60087/jaigs.v5i1.163

[16] Wang H, Cao P, Wang J, et al. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. Proceedings of the AAAI conference on artificial intelligence, 2022, 36(3): 2441-2449. https://doi.org/10.1609/aaai.v36i3.20144

[17] Ghosh S, Chaki A, Santosh K C. Improved U-Net architecture with VGG-16 for brain tumor segmentation. Physical and Engineering Sciences in Medicine, 2021, 44(3): 703-712. https://doi.org/10.1007/s13246-021-01019-w

[18] Meena S R, Soares L P, Grohmann C H, Van Westen C, Bhuyan K, Singh R P, Catani F. Landslide detection in the Himalayas using machine learning algorithms and U-Net. Landslides, 2022, 19(5): 1209-1229. https://doi.org/10.1007/s10346-022-01861-3

[19] Wang Y, Ye H, Cao F. A novel multi-discriminator deep network for image segmentation. Applied Intelligence, 2022, 52(1): 1092-1099. https://doi.org/10.1007/s10489-021-02427-x

[20] Prudviraj J, Vishnu C, Mohan C K. M-FFN: multi-scale feature fusion network for image captioning. Applied Intelligence, 2022, 52(13): 14711-14723. https://doi.org/10.1007/s10489-022-03463-x

[21] Liu J, Fan X, Jiang J, Liu R, Luo Z. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 105-119. https://doi.org/10.1109/TCSVT.2021.3056725

[22] Liu S, Fan H, Lin S, Wang Q, Ding N, Tang Y. Adaptive Learning Attention Network for Underwater Image Enhancement. IEEE Robotics and Automation Letters, 2022, 7(2): 5326-5333. https://doi.org/10.1109/LRA.2022.31561