

Underwater image enhancement via Transformer-based multi-scale spatial domain feature fusion network

HENG WU¹, JIAJIE HUA¹, FANLONG ZHANG^{1,2,*}, WEI MENG¹, CHUNHUA HE^{1,2,*}

¹ Guangdong Provincial Key Laboratory of Cyber-Physical System, School of Automation, Guangdong University of Technology, Guangzhou 510006, China

² School of Computer, Guangdong University of Technology, Guangzhou 510006, China

*Corresponding authors: zhangfanlong@gdut.edu.cn (FLZ), hechunhua@pku.edu.cn (CHH)

Underwater image enhancement (UIE) is widely used in underwater photography and vision fields to improve the contrast and clarity of underwater images. Due to problems of light scattering from suspended particles, water absorption, and refraction, underwater images often suffer color distortion and low contrast issues. To address these issues, we propose a UIE method with a Transformer-based multi-scale spatial domain feature fusion network (TMDFNet). We design a multi-scale convolution extraction module to fuse multi-scale and multi-level information and enhance the network adaptability on image feature extraction across diverse scales and levels. Simultaneously, we develop a spatial domain feature fusion module to improve the perceptual quality of underwater images and optimize the parameter configuration of the network. Moreover, we create a transformer module to capture intricate spatial relationships in underwater images, which enhances the network's comprehensive information extraction ability and the quality of reconstructed underwater images. A series of experiments are conducted to verify the effectiveness of TMDFNet. Quantitative and qualitative experiment results demonstrate that TMDFNet has competitive performance compared with other advanced UIE methods in subjective and objective evaluations.

Keywords: underwater imaging, image enhancement, multi-scale feature extraction, deep learning, Transformer.

1. Introduction

Underwater image enhancement (UIE) technology has garnered increasing attention in recent years owing to its wide-ranging applications in the field of underwater imaging. Optical imaging and computer vision technologies are commonly used in underwater studies and applications to capture sharp underwater images, with image enhancement being an indispensable aspect of this process [1]. UIE finds widespread applications in various domains, such as underwater archeology [2], marine scientific

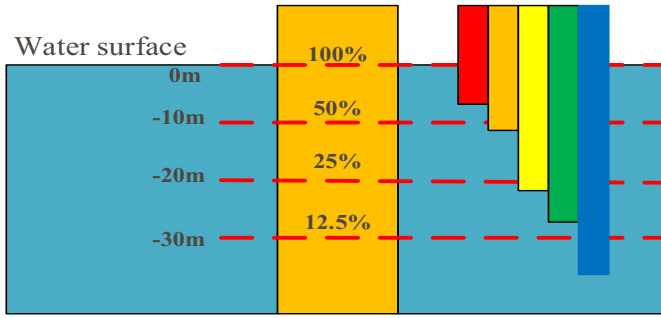


Fig. 1. Schematic of the light attenuation in water. The varying attenuation rates of light wavelengths (e.g., red, blue, and green) affect color balance and visibility in underwater images.

research [3,4], marine ecological research [5], underwater safety detection [6,7], and surveying, military, and defense, *etc.* Because of the impact of the aquatic environment, obtaining underwater images poses greater challenges compared to capturing images on land. As light penetrates beneath the surface of the water, it undergoes alterations due to phenomena like light attenuation and scattering, leading to issues such as chromatic aberration and substantial contrast reduction in the captured underwater images. As illustrated in Fig. 1, when light travels through water, it is absorbed by water, leading to the decrease of various wavelengths. Long wavelengths, such as red light, fade earlier than short wavelengths, like blue-green light [8]. Consequently, underwater images often exhibit a distinct blue or green color cast, and the lack of sufficient and comprehensive underwater image datasets further impedes the advancement of UIE techniques.

The challenges enumerated earlier have spurred the emergence of various UIE techniques, which can be broadly categorized into three groups: physical model-based [9,10], visual prior-based [11,12], and deep learning-based approaches [13,14]. By employing a particular predefined structure, physical model-based UIE techniques proficiently simulate the inverse operation of image degradation. Parameter estimation poses a challenge for these methods when adapting to dynamic underwater conditions. Yet, the primary obstacle for physical model-based UIE methods lies in their inability to unravel unknown physical phenomena, thereby restricting the applicability of existing models. Departing from dependence on physical models, visual prior-based UIE techniques reconstruct anticipated underwater images by adjusting pixel values, guided by considerations such as contrast, luminance, and saturation. As visual prior-driven UIE methods do not replicate physical processes, they may result in artificial coloring and overly aggressive enhancements. While deep learning-based methods [15,16] demonstrate their effectiveness in UIE, their performance relies significantly on the specific model architecture and the quality of the dataset at hand. Considering the substantial disparities between underwater imagery and terrestrial scenes, a convolutional neural network (CNN) that excels in terrestrial environments may face challenges when dealing with underwater conditions. Additionally, the insufficient presence of extensive under-

water datasets and the constrained diversity in underwater scenes hinder the effectiveness of deep learning-based methods. The challenge of capturing genuine underwater scene images persists.

Thus, due to the insufficiency of authentic underwater image datasets, many CNN-based methods train the networks using synthetic datasets. Nonetheless, the utilization of synthetic data for training presents challenges such as robustness, generalization, and effectiveness for CNN-based methods. Moreover, these methods face challenges in capturing global and long-range dependent features. Additionally, the convolutional kernels in CNN-based methods are inefficient in representing inconsistent features across different spatial and color channels, as well as in effectively extracting channel-dependent attenuation factors [17].

To effectively address the intricate degradation of underwater imagery, we propose TMDFNet, a comprehensive framework that synergizes spatial, frequency, and global contextual priors. By transitioning from conventional pixel-wise manipulation to a physics-aware optimization, TMDFNet systematically neutralizes wavelength-dependent attenuation and scattering effects. Specifically, we develop the multi-scale convolutional extraction module (MCEM) to facilitate adaptive spatial restoration. By utilizing parallel 1×1 , 3×3 , and 5×5 convolutional branches, the MCEM captures a diverse range of receptive field features. This design, integrated with channel attention and dynamic weight coefficients, allows the network to adaptively blend local chromatic details with global statistical priors, thereby mitigating the blurring and color wash-out caused by complex light scattering. To target the physical root of color distortion, the spatial domain feature fusion module (SDFFM) introduces frequency-domain spectral reconstruction. By leveraging the Fast Fourier Transform to decouple amplitude and phase spectra, the module specifically reconstructs the amplitude components that govern color and energy distribution. This approach directly rectifies wavelength-selective absorption by compensating for the non-uniform spectral decay across the R, G, and B channels, ensuring high-fidelity color reproduction. Finally, the transformer module (TM) is incorporated to ensure global chromatic harmony and visual consistency. Utilizing a linear attention mechanism to establish long-range spatial dependencies, the TM captures ambient illumination patterns and global color variances across the entire scene. This facilitates fine-grained channel mapping via a multilayer perceptron (MLP), which systematically balances the light field and enhances the overall chromatic stability of the restored image. The TMDFNet is trained on the real-world underwater image benchmark (UIEB), comprising 890 images that depict diverse underwater scenes. These scenes include underwater photos with coral formations, individuals swimming underwater, and underwater rock formation images. The effectiveness of TMDFNet is validated by many experiments on the UIEB dataset.

The main contributions of this work are described as follows.

(1) We propose a TMDFNet for UIE. Unlike most existing fusion methods, TMDFNet can capture information at various scales in underwater images, enhancing the ability of the model to express features at different levels. This capability contributes to the improvement of feature extraction.

(2) We design an MCEM to extract information at different scales and levels. Additionally, as most pixels in local neighborhoods of high-resolution feature maps in underwater images share similar features, performing pairwise attention calculations between all pixels is inefficient and redundant. To address this issue, we design an SDDFM to calculate attention within local neighborhoods of pixels, thereby improving computational efficiency and feature fusion.

(3) Due to the influence of factors such as scattering, absorption, and background lighting, underwater images are prone to noise and blur. To solve these issues, we propose a TM that utilizes self-attention and linear attention mechanisms to capture complex spatial relationships in images and improve image quality.

(4) Extensive experiments indicate that TMDFNet possesses outstanding performance and optimal generalization ability, achieving superior results on the real-world underwater image dataset UIEB compared with other state-of-the-art (SOTA) methods. The proposed TMDFNet can find applications in underwater photography and vision to enhance the sharpness, contrast, and overall quality of underwater images

2. Related works

2.1. Underwater image enhancement

Various methods have been developed for UIE, which can be classified into two main groups: conventional approaches and deep learning methods. When conventional UIE techniques are utilized for underwater images, they suffer the problems of light absorption and scattering. Specifically, they face difficulties in accurately correcting color distortions, which can lead to artificially vivid colors in the processed images. Moreover, traditional approaches depend on modifying pixel values to improve underwater image quality, but their adaptability to diverse underwater scenarios is limited. This limitation presents a challenge for traditional techniques to effectively grasp and adjust to the nuanced characteristics inherent in underwater imagery, leading to issues like color bias and artifacts in the resultant images. LIU *et al.* [18] suggested a simulation method based on models for underwater image generation and applied an enhancement approach based on learning to tackle various degradation issues observed in underwater images. Based on [18], LIU *et al.* [19] expanded the image formation model (IFM), assuming locally consistent illumination and accounting for non-uniform light sources. They introduced a novel model aimed at addressing degradation in underwater images resulting from both water conditions and illumination factors. Unlike [19], MARQUES *et al.* [20] employed the dark channel prior (DCP) to enhance low-light underwater images. Their utilization of DCP aimed at enhancing the brightness and contrast of poorly lit areas, while they also introduced a contrast-guided methodology for calculating dark channels and transmission maps. LIU *et al.* [21] developed a multi-purpose dehazing framework for nighttime haze images to cope with multiple adverse degradation factors of nighttime haze images. WU *et al.* [22] proposed a visual transformer (ViT) based network (called WaterFormer) to improve the quality of underwater images.

Deep learning methods for UIE can extract diverse feature representations from low-quality underwater images and subsequently utilize these features to accomplish high-quality restoration of underwater images. TANG *et al.* [23] proposed a conditional generative adversarial framework utilizing an attention UNet architecture for real-world UIE. This model employs an attention mechanism to eliminate extraneous information, consequently enhancing the quality of images. HUANG *et al.* [24] introduced a UIE technique employing an adaptive group attention-based multi-scale cascade transformer. This transformer dynamically selects visually complementary channels and reduces the number of attention parameters. QI *et al.* [25] reported a UIE network incorporating semantic information for enhanced feature learning within regions. Additionally, they introduced a semantic region enhancement module designed to learn multi-scale perceptual regions and augment local features. TIAN *et al.* [26] presented a UIE network that integrates heterogeneous feature fusion and dynamic enhancement. This network extracts both low-level and high-level features, facilitating improved communication among modules and preserving a more comprehensive range of image details. However, it is worth noting that most deep learning-based methods mainly rely on spatial domain processing and often ignore the rich global information in the Fourier domain. Therefore, innovative solutions are needed. To solve this problem, a method based on spatial domain feature fusion is proposed to utilize the rich information in the Fourier domain.

2.2. Multi-scale feature fusion

Single-scale image enhancement methods have difficulty in capturing both detail information and global features at different scales, which limits their overall effectiveness. In addition, single-scale strategies are usually insufficient to cope with the diversity and complexity of underwater images. In this context, multi-scale feature fusion (MFF) plays a vital role in underwater image enhancement (UIE) as it enables the network to effectively handle both fine details and global structures, thus significantly improving image quality. Moreover, MFF enhances the adaptability of the network, enabling it to handle objects and structures at different scales and enhancing its performance in diverse underwater environments and scenarios. By incorporating multi-scale features, the network acquires a more comprehensive and enriched representation of image features, effectively capturing the complex textures and structures presented in underwater scenes. MI *et al.* [27] introduced a UIE technique, which initially conducts white balance adjustments on the deteriorated underwater image to eliminate color shifts and regain its authentic appearance. Subsequently, a multi-scale gradient-domain contrast enhancement method is employed to enhance visibility while compensating for the reduction in color saturation based on the estimated transmittance. FAN *et al.* [28] demonstrated a UIE algorithm that merges color correction with multi-scale fusion. Initially, an improved multi-scale Retinex algorithm is applied for color correction, followed by the enhancement of image edge details through contrast-limited adaptive histogram equalization and sharpening algorithms. Lastly, employing the Laplacian pyramid and

Gaussian pyramid for multi-scale decomposition and layer-by-layer fusion reconstruction achieves a comprehensive enhancement of underwater images. CHEN *et al.* [29] proposed an underwater scene image enhancement method based on the MFF network (MFFN). To extract multi-scale features, they designed modules for fused feature extraction, feature fusion, and attention reconstruction. These measures not only enhance the adaptability and visual effects of the scene but also introduce multiple objective functions for supervised training to match nonlinear mappings. UEKI *et al.* [30] developed a UIE method that combines residual feature attention blocks with a multi-scale, multi-patch structure. The polyhedron network is used to extract local features to adapt to various underwater images, which are typically non-uniform. QIAO *et al.* [31] proposed a high-performance UIE model. They introduced a new attention mechanism and an MFF method, effectively combining high-dimensional and low-dimensional elements, enhancing salient features, and suppressing less critical components. However, the existing multi-scale fusion network is sensitive to noise and it is difficult to effectively maintain the boundaries and details of the image. To this end, this paper proposes a method that combines the attention mechanism to enhance the key features in the image and suppress irrelevant information such as noise, thereby improving the contrast of underwater images.

2.3. Transformer methods

Inspired by the success of Transformers in NLP and high-level vision tasks, they have been applied to image restoration, outperforming previous CNN-based methods by effectively modelling non-local information. YANG *et al.* [32] illustrated a Transformer-based approach for UIE and depth estimation. This method leverages the benefits of joint learning for both enhancing underwater images and estimating depth simultaneously. LIU *et al.* [33] proposed a UIE model based on vision transformers and CNN. This model incorporates MobileVitV2 block fusion augmentors and global-local feature fusion block to learn underwater feature representation, achieving success in feature fusion and global-local feature extraction. BADRAN *et al.* [17] utilized a cycleGan network to generate a new underwater dataset from existing datasets. Additionally, they employed the Uformer structure for model training and incorporated a depth estimation module to learn object depths. REN *et al.* [34] proposed an enhanced swin-convs transformer for super-resolution (ERSCT-SESR) method, which is based on U-Net. In this approach, the Swin Transformer is integrated into the U-Net architecture to enhance its capacity for capturing global dependencies. WU *et al.* [35] presented a cooperative enhancement watermarking transformer (CEWformer) for concurrent UIE and watermarking. They employed channel self-attention transformers (CSAT) to improve channels experiencing significant color degradation by leveraging channel correlations. The traditional transformer structure has high computational complexity and fails to capture local details. The transformer we designed removes redundant information and consists of only simple MHSA and FFN. In addition, a convolution module is designed to capture local features, thereby enhancing the contrast of underwater images.

3. Proposed method

3.1. Underwater image model

Figure 2 shows the principle of the underwater camera imaging model. After the light reflects from the underwater target, it is focused through the lens on the sensor of the underwater camera to form an image. Generally, the lights in the underwater imaging model can be expressed as a linear combination of three components [36]: direct component (DC), forward scattering component (FSC), and backward scatter component (BSC). When the light passes through the water surface, it is focused on the underwater target after reflection and transmission. The underwater camera sensor receives lights to form an image. During this process, when the light interacts with suspended particles in the water, the light is scattered in a random direction, which is called the FSC, resulting in a reduction in the brightness and contrast of the image. Besides, water absorption also reduces the brightness and contrast of the image. Additionally, the scattered light produced by suspended particles in the BSC [37] causes image quality degradation and blurred details. Due to the influence of FSC and BSC, the imaging sensor of the underwater camera receives much interference, causing color distortion and low contrast issues and degrading the quality of the captured images.

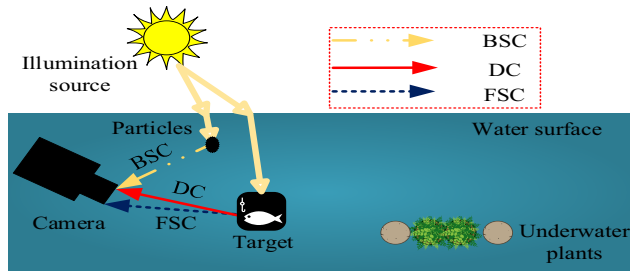


Fig. 2. Sketch map of an underwater imaging model. BSC, backward scatter component; DC, direct component; FSC, forward scattering component.

3.2. Overall network architecture

To solve color distortion and low contrast issues in underwater images and challenges in previous CNN-based approaches, a TMDFNet is proposed. Figure 3 shows the overall architecture of the proposed TMDFNet. TMDFNet includes five parts. This design aims to extract, encode, compress, decode, and reconstruct the underwater images. Firstly, the underwater image is input into a feature extraction (FE) part to extract the image features directly. Secondly, a feature encoding (FED) part is designed to map the input feature (the channel number is $N = 3$) into a deeper feature space, where the channels are increased to $N = 64$ and spatial dimensions (height and width) are reduced. Thirdly, a feature compression (FC) part located in the middle of the network learns valuable feature compression while preserving the dimensionality of the feature space. Fourthly, the feature decoding (FDD) part further processes the features with differ-

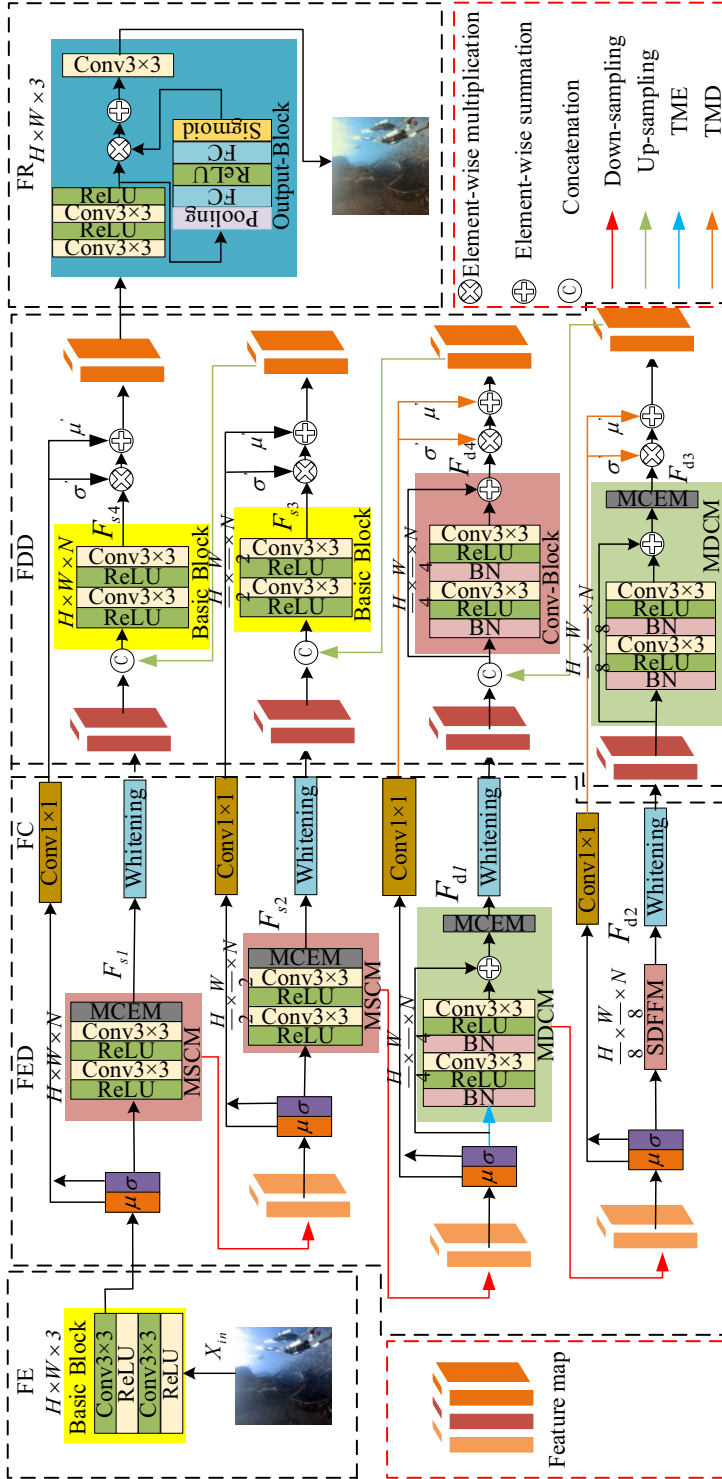


Fig. 3. Structure diagram of TMDFNet. FE, feature extraction; FED, feature decoding; FDD, feature decoding; FR, feature reconstruction; TME, transformer encoder; TMD, transformer decoder; ReLU, rectified linear unit; $3 \times 3 \text{Conv}$, convolution with a 3×3 kernel; BN, batch normalization.

ent dimensions, where the spatial dimensions (height and width) are increased from $H/8 \times W/8 \times N$ to $H \times W \times N$. The channel number in the FED and FDD part remains unchanged, where N is equal to 64. Finally, the feature reconstruction (FR) part outputs the UIE images. In Fig. 3, μ and σ are the mean and standard deviation tensors, respectively, which are obtained by the positional normalization (PONO) operation [38] on the output feature data of the previous layer. The PONO operation can help the network to capture rough structural information. After the PONO operation, the mean and standard deviation tensors (μ and σ) go into the FC part and are processed by two separate 1×1 convolutions, generating corresponding feature maps A and B. Feature maps A and B are further processed through two 1×1 convolutions and tensors μ' and σ' are generated. Here, the tensors μ' and σ' are used for subsequent feature processing, enabling feature scaling and shifting operations.

Assume that an original underwater image $X_{in} \in \mathbb{R}^{H \times W \times 3}$ is input into the network, where H and W are the height and width of the image, respectively, and 3 is the channel number of the input image. Initially, the image undergoes the FE part, which is designed to learn the fundamental features of the image and perform simple information extraction.

Subsequently, the image passes through the FED part. In this part, we propose a flexible multi-scale shallow convolution module (MSCM) to extract information at different scales and levels, enhancing the model's ability to understand and process contextual information in underwater environments. This process yields shallow embedded features $F_{s2} \in \mathbb{R}^{H/2 \times W/2 \times N}$ with N channels. These shallowly embedded features are input into the proposed transformer encoder (TME) to capture global information and improve the generalization ability of the model. Next, the shallow features extracted from the TME are input into the multi-scale depth convolution module (MDCM), which utilizes contextual information to fuse features to generate deeply embedded features $F_{d1} \in \mathbb{R}^{H/4 \times W/4 \times N}$.

Therefore, these features are transformed into deep-level parts. To address the current issue where most methods for restoring degraded images heavily rely on spatial domain processing, and traditional convolution methods often neglect the rich global information in the Fourier domain, we propose an SDFFM (as shown in Fig. 5), whose output is the deep feature map $F_{d2} \in \mathbb{R}^{H/8 \times W/8 \times N}$.

The specific principle of the whitening operations [39] involves transforming the input data to eliminate redundant information. This process aims to reduce the correlation between adjacent pixels and ensure that all features exhibit the same mean and variance. Thus, the whitening operations are introduced in the FC part to reduce redundancy among image features, decrease the correlation between each component, and equalize their variances. To mitigate the impact of degradation diversity and reinforce the connection between the FED and FDD, four 1×1 convolutions are incorporated at the skip connection line. These convolutions help to enhance global dependencies. After passing through the bottom two feature extraction modules (SDFFM and MDCM) in the FED and FDD parts, the output feature maps are recorded as $F_{d3} \in \mathbb{R}^{H/64 \times W/64 \times N}$.

Here, the output feature map size remains unchanged, and the whitening operation eliminates redundant information in the input data, which improves the efficiency of core information extraction.

Like the FED part, TMD is designed on the skip connection lines of the FDD part, as the orange line shown in Fig. 3. By weighing global and local information in different frequency domains, TMD improves the retention of the overall quality and local details of underwater images and increases the adaptability of the proposed network to complex underwater environments. After passing through four reconstruction modules (MDCM, Conv-Block, and two BasicBlocks), the shallow features are recorded as $F_{s4} \in \mathbb{R}^{H \times W \times N}$. Finally, after the feature maps $X_{\text{out}} \in \mathbb{R}^{H \times W \times 3}$ pass through the FR part, a clear underwater image is restored by projecting the embedded features onto the RGB channels.

In the underwater environment, the light of different wavelengths experiences different attenuation rates, which results in significant differences among the RGB channels in underwater images, leading to reduced contrast and color distortion of the image. In the proposed model, we design MSCM and MDCM. Existing modules only enhance underwater images from a single angle, while the designed MCEM enhances the receptive field of the network from a multi-scale perspective and improves the robustness and adaptability of the network by capturing features at different scales. In addition, previous image enhancement modules mainly rely on spatial domain processing and frequency domain information is not utilized. The proposed SDDFM makes full use of frequency domain information through the Fourier domain to estimate and adjust the illumination distribution in the image, thereby improving the contrast and clarity of the image. However, when the input image contains a lot of noise or distortion, the frequency domain information is lost, affecting the enhancement effect. Therefore, MDCM, TME, and ACM are introduced to capture local details and remove noise, further improving the image enhancement effect. Additionally, underwater images are susceptible to noise and blur, and traditional transformers are limited. Therefore, we improve the Transformer module to leverage its efficient parallel computing power to remove noise and blur in images.

3.3. Multi-scale convolution extraction module

In underwater environments, the non-uniform attenuation of light wavelengths—characterized by the rapid disappearance of red light and the relative persistence of blue-green frequencies—induces a severe spectral imbalance. This physical phenomenon leads to a compounding effect of chromatic distortion and contrast suppression, creating a spatially variant restoration challenge that remains largely unresolved by conventional methods. To solve the contrast reduction and color distortion in underwater images, an MCEM that can enhance the receptive field of the network is designed. As illustrated in Fig. 4, given an input feature map x , three feature maps $X_{i=1,2,3}$ are gen-

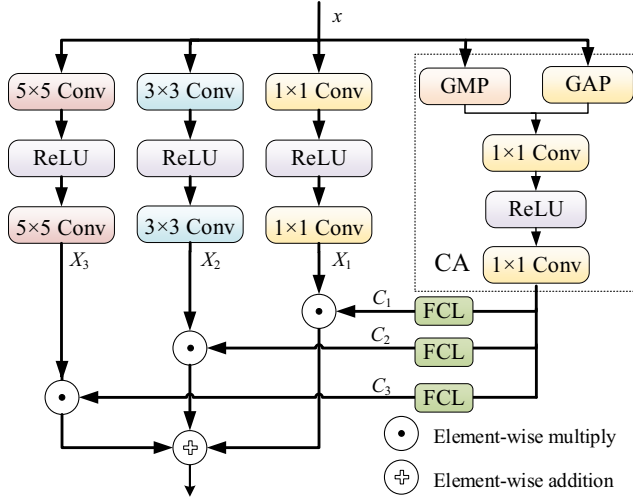


Fig. 4. Structure diagram of MCEM. CA, channel attention; FCL, fully connected layer.

erated by three convolutional blocks with different kernel sizes of 1×1 , 3×3 , and 5×5 , which are expressed as,

$$X_1 = \text{Conv5_5}(x) \quad (1a)$$

$$X_2 = \text{Conv3_3}(x) \quad (1b)$$

$$X_3 = \text{Conv1_1}(x) \quad (1c)$$

where $\text{ConvZ_Z}(\cdot) = \text{ConvZ}(\text{ReLU}(\text{ConvZ}(\cdot)))$ denotes the module composed of a convolution operation $\text{ConvZ}(\cdot)$, a linear rectification function $\text{ReLU}(\cdot)$, and a convolution operation $\text{ConvZ}(\cdot)$. $\text{ConvZ}(\cdot)$ denotes the convolution operation with a kernel size of $Z'Z$, and $Z = 1, 3, 5$. By automatically extracting and fusing the main features in the feature maps of three different scales, the structure helps the network to better adapt to information of different scales and levels, improving the performance of the model. Additionally, the input feature is also processed by the channel attention (CA) block [40], and the output feature map X_{att} is expressed as,

$$X_{\text{att}} = \text{CA}(x) \quad (2)$$

where $\text{CA}(\cdot)$ represents the channel attention function. By introducing channel attention, MCEM can reduce sensitivity to noise or irrelevant features while enhancing the degree of attention between different channels in the input feature map x . This helps the network capture key features in images more efficiently. After CA, we combine

the feature map with a fully connected layer (FCL), whose output feature map is represented as,

$$C_1 = \text{FCL}(X_{\text{att}}) \quad (3a)$$

$$C_2 = \text{FCL}(X_{\text{att}}) \quad (3b)$$

$$C_3 = \text{FCL}(X_{\text{att}}) \quad (3c)$$

where $\text{FCL}(\cdot)$ denotes a fully connected layer which contains two linear sub-layers. The function $\text{FCL}(\cdot)$ is to integrate information and represent features of the feature map output by the channel attention mechanism (CA), as well as improve the model's ability to focus on and extract important features. Finally, the above features are summed by an addition operation to obtain the output feature map, which is expressed as,

$$Y = C_1 \times X_1 + C_2 \times X_2 + C_3 \times X_3 \quad (4)$$

where $\times$ signifies the element-wise multiply and $+$ denotes the addition of feature maps. In Eq. (4), the feature maps and ($Z = 1, 3, 5$) have the same spatial dimensions. By element-wise multiplication of their corresponding elements, the module can learn relationships between different branches, thereby enhancing the representative capacity of the network. Subsequently, feature maps from different branches are combined by the addition operation to achieve feature fusion at different scales.

3.4. Spatial domain feature fusion module

In underwater images, global background illumination and textures can be decomposed in the Fourier domain [41]. However, current methods for restoring degraded images predominantly rely on spatial domain processing, often neglecting the abundant global information (the frequency components) in the Fourier domain. Specifically, the Fourier transform reveals the existence and intensity of different frequencies in an image, reflecting the overall structure and characteristics of the image. The spatial transform is directly related to the pixel values of the image, showing specific details, colors, and textures. In underwater image enhancement, the global information in the Fourier domain can help identify and suppress the scattering and absorption effects of water, thereby improving the visibility of the image, while the spatial domain information is mainly used for detail recovery. To solve the color distortion in underwater images, as depicted in Fig. 5, we propose an SDDFM that uses the Fourier domain to help estimate and adjust the illumination distribution in the image, thereby improving the contrast and clarity of the image and suppressing the impact of uneven illumination.

Initially, input features undergo processing through an MDCM and an attention convolution module (ACM). MDCM extracts information from input features through shallow operations and fuse features at different levels to improve the expressive ability of the model. MCEM is used to fuse information from multiple levels of features and capture input feature information from multiple dimensions. ACM is used to sup-

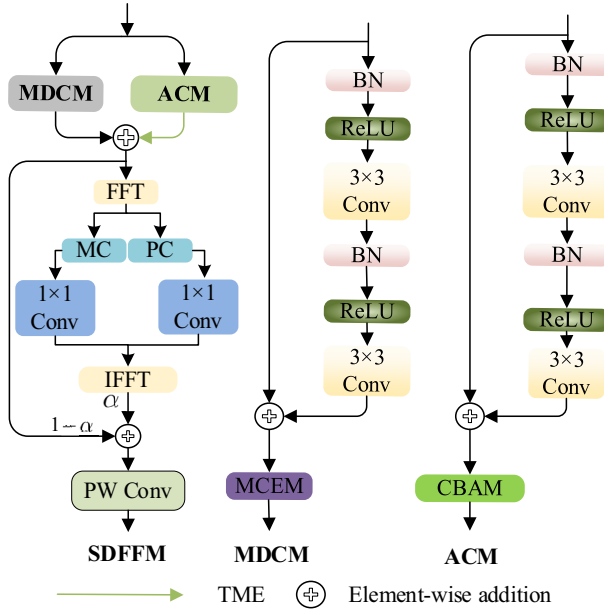


Fig. 5. Structure of SDDFM. MC, magnitude component; PC, phase component; TME, transformer encoder; PW Conv, Point-wise convolution.

press irrelevant information and highlight important features by introducing channel attention (CBAM), which can finely allocate and process input features, focusing on the most important features. Thus, the image details can be better extracted by summing the features output by MDCM and ACM. Moreover, TME is employed to capture information bias from the input features, improve the generalization capability of the network, and reduce overfitting to the training data. In SDDFM, the FFT is utilized to transform spatial domain feature maps into the frequency domain, which is expressed as magnitude components (MC) and phase components (PC), respectively. The 1×1 convolutions are utilized to eliminate noise and irrelevant information in the feature maps, thereby achieving image enhancement. Conversely, IFFT transforms the feature maps from the frequency domain to the spatial domain for the feature map reconstruction. FFT and IFFT operations help the network extract global features and better capture the frequency domain information of the image, thus achieving effective image enhancement.

The input features are $F_{d2} \in \mathbb{R}^{HW/64 \times N}$ and the critical information of the SDDFM is expressed as,

$$F' = MC(F_{d2}) + CC(TF(F_{d2})) \quad (5)$$

$$F_{MAG}, F_{PHA} = f_{FFT}(F') \quad (6)$$

$$F_O = (1 - \alpha)F' + \alpha \left[f_{IFFT}(f_{FC}(F_{MAG}), f_{FC}(F_{PHA})) \right] \quad (7)$$

where $MC(\cdot)$, $CC(\cdot)$ and $TF(\cdot)$ denote MDCM, ACM, and TM, respectively. Additionally, F_{MAG} and F_{PAH} specifically denote the magnitude component and phase component of the features, respectively. $f_{IFFT}(\cdot)$ is the Fourier domain convolution operation and $f_{FFT}(\cdot)$ is the inverse fast Fourier domain transform. The hyperparameter α in SDFFM is a constant that controls the fusion ratio of the spatial frequency domain information, enhancing the perceptual field and quality of the network.

3.5. Transformer module

Due to the serious color deviation caused by the absorption and scattering of light in water and the large-scale blur and diffusion effects in underwater environments, underwater images are prone to noise and blur. The transformer [42] structure can perform global information fusion and reconstruction, establish long-range feature dependencies, and its efficient parallel capabilities can cope with the above challenges. To solve the noise interference in underwater images, as illustrated in Fig. 6, a TM is proposed

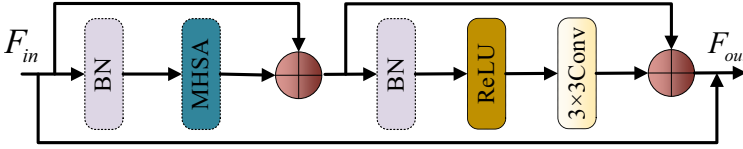


Fig. 6. Structure of TM. $\oplus$ is the element-wise addition.

to leverage the multi-head self-attention (MHSA) mechanism to capture complex spatial relationships in images, thereby enhancing image quality. The TM allows the model to jointly infer attention from different representation subspaces, connecting multiple attention heads and further examining them through a feedforward network. Here, three heads are used in the MHSA mechanism. Assuming that the feature map size of the input feature map is $F_{in} \in \mathbb{R}^{C \times H \times W}$. The input feature map is enhanced by the TM by capturing important information in the image via the attention mechanism. Residual connections are employed to preserve information of the original image. Subsequently, the feature maps are processed by a convolution layer to obtain enhanced output feature map. The feature map extraction process of TM is expressed as,

$$F_{out} = F_{in} + \text{Conv}_{3 \times 3}(f_{\text{relu}}(\text{BN}(F_{in} + \text{MHSA}_2(\text{BN}(F_{in})))) \quad (8)$$

$$F_{out} = F_{in} + \text{Conv}_{3 \times 3}(f_{\text{relu}}(\text{BN}(F_{in} + \text{MHSA}_3(\text{BN}(F_{in})))) \quad (9)$$

where $\text{BN}(\cdot)$, $f_{\text{relu}}(\cdot)$, and $\text{Conv}_{3 \times 3}(\cdot)$ signify the batch normalization, ReLU, and 3×3 convolution, respectively. $\text{MHSA}_x(\cdot)$ denotes the multi-head attention with a dimension of x , and $x = 2, 3$. Note that Eq. (7) and Eq. (8) are used in the TME and TMD, respectively.

3.6. Loss function and training process

The mean square error (MSE) and perceptual similarity are used to train TMDFNet with a given training set $\{Y_{\text{raw}}^i, Y_{\text{clear}}^i\}_{i=1}^I$, where Y_{raw} is the original underwater image, Y_{clear} is the clean underwater image, and I is the total number of image pairs in the training set. The overall loss function $\wp_{\text{all}}$ includes two parts, which are defined as,

$$\wp_{\text{all}} = \wp_{\text{MSE}} + \lambda \wp_{\text{PL}} \quad (10)$$

where λ is the weight.

MSE is calculated as the square of the pixel-wise difference, and the loss function $\wp_{\text{MSE}}$ is defined as,

$$\wp_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n (Y_{\text{clear}}^i - \hat{Y}_{\text{clear}}^i)^2 \quad (11)$$

where $\hat{Y}_{\text{clear}}^i$ is the output enhanced underwater images and n is the pixel numbers of each image. The perceptual similarity type focuses on the high-level semantic information of the image rather than the pixel-level differences, which defines the perceptual difference between the enhanced image and the clean image. Thereby, the loss function $\wp_{\text{PL}}$ is defined as,

$$\wp_{\text{PL}} = \frac{1}{n} \sum_{i=1}^n \left[\varphi_{mn}(Y_{\text{clear}}^i) - \varphi_{mn}(\hat{Y}_{\text{clear}}^i) \right]^2 \quad (12)$$

where $\varphi_{mn}(\cdot)$ denotes the implicit function of the VGG16 model.

4. Experiments on public dataset

4.1. Training details

The real-world UIEB [43] dataset is utilized for training and testing of the proposed model. The UIEB dataset comprises 890 images, where the initial 700 images serve as the training set and the remaining ones serve as the test set. Besides, the Real-world datasets LSUI [43], EUVP [44], UFO-120 [45], and RUIE [46] are used for testing. The construction of the test dataset is detailed as follows:

- Test-L900: 900 images from the LSUI [43];
- Test-E512: 512 images from the EUVP [44];
- Test-U120: 120 images from the UFO-120 [45];
- Test-R100: 100 images from the RUIE [46].

Note that Test-R100 only contains original images and has no real underwater images. The images in the test dataset (*e.g.*, Test-L900, Test-E512, *etc.*) are randomly selected from the original dataset (*e.g.*, LSUI, EUVP, *etc.*).

During training, all images are resized to a fixed dimension of 256×256 pixels, and the test images are all adjusted to 256×256 pixels. TMDFNet is implemented using Python and the PyTorch framework on the NVIDIA RTX 3070 GPU. The training process iterations are set as 256, and we utilize the Adam optimizer with a batch size set to 4 and an initial learning rate of $1e-4$ during training. Additionally, both FED and FDD parts in TMDFNet are comprised of four layers, where three downsampling and three upsampling operations are integrated sequentially between them.

4.2. Evaluation metrics

For a comprehensive and fair evaluation, we adopt three widely used reference metrics: peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [47], learning perceptual image patch similarity (LPIPS) [48], and two reference-free metrics: the universal image quality metric (UIQM) [49] and Twice Mixing [50]. PSNR serves as a metric for assessing image reconstruction quality and quantifying the accuracy of image reconstruction. SSIM functions as a metric for evaluating image similarity, considering the structural, luminance, and contrast differences between the original and reconstructed images. LPIPS is a deep learning-based metric for image quality assessment, aligning more closely with human perception of image quality. The UIQM evaluation metric uses the underwater image colourfulness measure (UICM), the underwater image sharpness measure (UISM), and the underwater image contrast measure (UIConM) as the evaluation basis and is expressed as a linear combination of these three. When the value of UIQM is larger, the color balance, clarity, and contrast of the image are better. The core idea of Twice Mixing is to randomly generate two mixing ratios for generating training examples and corresponding quality rankings. Unlike typical linear combination mixing algorithms, Twice Mixing considers the non-uniformity and non-linearity of the human visual system and evaluates the quality of two virtual instances through random mixing ratios. The larger the Twice Mixing, the better the color balance, clarity, and contrast of the image. A better UIE method should have higher PSNR, SSIM, UIQM and Twice Mixing and lower LPIPS. Parameters (Params) represent the number of trainable parameters in a model, affecting the complexity and capacity of the model. FLOPs denote the floating-point operations, reflecting the computational cost and efficiency of a model. Both parameters and FLOPs are used to measure the performance and resource requirements of a model.

4.3. Comparison with other SOTA methods on the UIEB dataset

The proposed method is evaluated using 190 reference images from the UIEB dataset [43]. To demonstrate the superiority of TMDFNet, seven SOTA deep learning UIE methods are selected for comparison, including U-shape [17], PUIE-net [51], STSC [52], UWnet [53], FiveAPlus [41], URSCT [34], UIE-UnFold [54], and SCnet [39].

Figure 7 shows the experimental results of the TMDFNet and comparison SOTA methods. UWnet darkens colors, resulting in reduced brightness and contrast in the restored image. Although PUIE-net and Ushape produce brighter images than UWnet, their im-

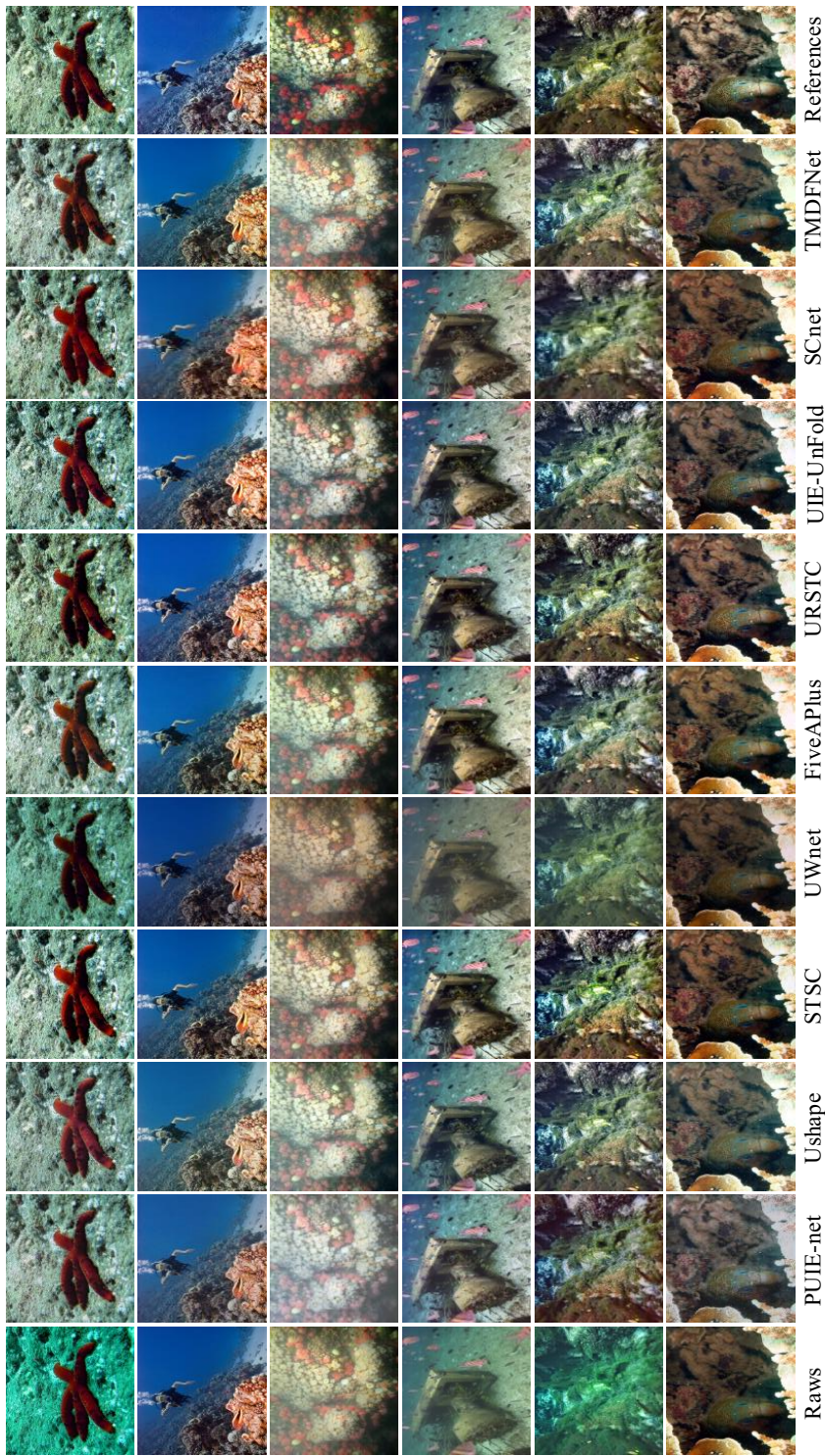


Fig. 7. The results of different methods on the UIEB.

age contrast and brightness are lower. FiveAPlus and UIE-UnFold improve the clarity and brightness of the image, but it causes an overexposure effect, producing a white halo at the edge of the scene. STSC and URSTC can improve the quality of the original image and handle most cases of color deviation. However, STSC and URSTC cannot process the local information in the original image, leading to a loss of details in the recovered image. SCnet introduces significant noise in enhanced underwater images. The above methods only consider the local information of the image from the perspective of a single scale and ignore the importance of global information in the underwater image. Therefore, they cannot fully capture the various details and structural information of the image. These methods may have limitations when processing underwater images, especially when faced with phenomena such as light attenuation and scattering. On the contrary, TMDFNet adopts a multi-scale context fusion strategy, allowing the network to consider information at different scales simultaneously and understand the content of underwater images more comprehensively. With this strategy, TMDFNet can more accurately control texture details, thereby restoring and enhancing overall degraded underwater images.

As shown in Fig. 8, UWnet cannot intuitively depict the underwater recovery effect, and the image has color distortion and poor edge detail processing. STSC lacks precise color control, resulting in noticeable color distortion. SCnet not only lacks color control but also exhibits many artifacts. Like, URSTC has poor control over the underwater color and cannot eliminate the green tint caused by the scattering of light, showing a wired ability to enhance vibrant colors. PUIE-net and Ushape improve the clarity and brightness of the image, but the overall image is overexposed, resulting in the loss of image details. Although FiveAPlus and UIE-UnFold restore images with high contrast and provide precise color control, it compromises the preservation of local details during the enhancement process. The above methods only restore the image color from a single-scale perspective and mainly optimize the local area color information of the image but ignore the global color information of the image. This local area optimization may generate enhanced images that are locally too bright or contrasty but lack overall consistency and balance. In contrast, due to the adoption of a multi-scale context fusion strategy, TMDFNet comprehensively considers the local features and global context of the image and achieves adaptive adjustment. This adaptability helps TMDFNet better adapt to changes such as different depths, lighting conditions, and water quality and improves robustness and versatility during the UIE process. Additionally, TMDFNet can control color and alleviate color distortion caused by light scattering in underwater environments, which provides a good perception effect for contrast and texture details.

Besides, we utilize objective evaluation metrics such as PSNR, SSIM, LPIPS, and UIQM to objectively assess the superiority of the proposed network. As shown in Table 1, the proposed TMDFNet outperforms previous UIE methods on four evaluation metrics. Particularly, compared to SCnet, TMDFNet improves by 1.869 dB in PSNR, 0.0887 in SSIM, and slightly decreases in LPIPS by 0.1679. Furthermore,



Fig. 8. The results of different methods on the UIEB.

Table 1. Performance comparisons on the UIEB dataset. Bold means the best, and underlined denotes the second best. The symbols $\uparrow$ and $\downarrow$ indicate a higher score is better and a lower score is better, respectively.

Methods	U-shape	PUIE-net	STSC	UWnet	FiveAPlus	URSCT	UIE-UnFold	SCnet	TMDFNet
PSNR $\uparrow$	20.1186	20.8532	19.3779	16.7796	<u>21.9978</u>	21.7813	21.9380	20.1793	22.0483
SSIM $\uparrow$	0.7370	0.8237	0.6791	0.6843	0.8337	<u>0.8804</u>	0.8867	0.8037	0.8924
LPIPS $\downarrow$	0.2180	0.1596	0.1501	0.4902	0.1610	<u>0.1253</u>	0.1265	0.2847	0.1168
UIQM $\uparrow$	1.931	1.732	<u>1.999</u>	1.5250	1.947	1.967	1.1538	1.783	2.012
Twice Mixing $\uparrow$	1.5724	0.9033	1.0672	1.4023	2.1552	1.8967	<u>2.3228</u>	<u>0.9171</u>	2.4408
Flops (G) $\uparrow$	2.983	30.093	186.608	21.670	18.740	18.105	112.7846	23.555	<u>48.766</u>
Params (M) $\downarrow$	22.817	1.401	25.288	0.2190	0.008	11.261	<u>4.1873</u>	<u>0.817</u>	1.454
Training time (s) $\downarrow$	35.3013	12.9915	31.2867	<u>5.0934</u>	0.0977	62.6617	52.2353	12.6181	61.3375
FPS $\uparrow$	28.3275	76.9731	31.9623	196.3292	10.2304	15.9586	<u>19.1441</u>	<u>79.2509</u>	16.3032

TMDFNet is the best on reference-free metric UIQM. As shown in Table 1, TMDFNet has advantages in terms of efficiency and performance. Specifically, compared with SOTA methods, TMDFNet has moderate Params and FLOPs. TMDFNet reduces the consumption of computing resources while maintaining high image enhancement performance. This moderate number of parameters and FLOPs makes TMDFNet more practical and efficient in resource-constrained environments. In addition, this balance means that TMDFNet can provide better model performance without significantly sacrificing image quality, making it a good balance between computational efficiency and image enhancement effect. In addition, the training time and FPS of proposed model are 61.3375 and 16.3032, indicating that the transformer computational load is too high to be used in resource-constrained environments. In future work, we will try to reduce the number of parameters in the model by applying model pruning to remove unnecessary neurons or convolutional kernels, thereby optimizing the training time and FPS of TMDFNet. Objective and subjective evaluations show that TMDFNet has achieved sound restoration effects in improving image clarity and color saturation and is more in line with human visual perception.

4.4. Comparison results with other SOTA methods on EUVP, UFO, LSUI and RELU dataset

4.4.1. Qualitative comparison results

The experimental results of the four datasets, Test-L900, Test-E512, Test-U120, and Test-R100, are shown in Figs. 9–12.

Figure 9 shows the experimental results on the Test-E512. UWNet produces significant noise and color deviation in the reconstructed images. FiveAPlus and Ushape sharpen the edges of fish excessively in restored images. Although URSCT and UIE-UnFold corrects color deviations, image details are very blurry. Compared with UWNet, UShape and PUIE-net eliminate color bias, but the image contrast is significantly re-

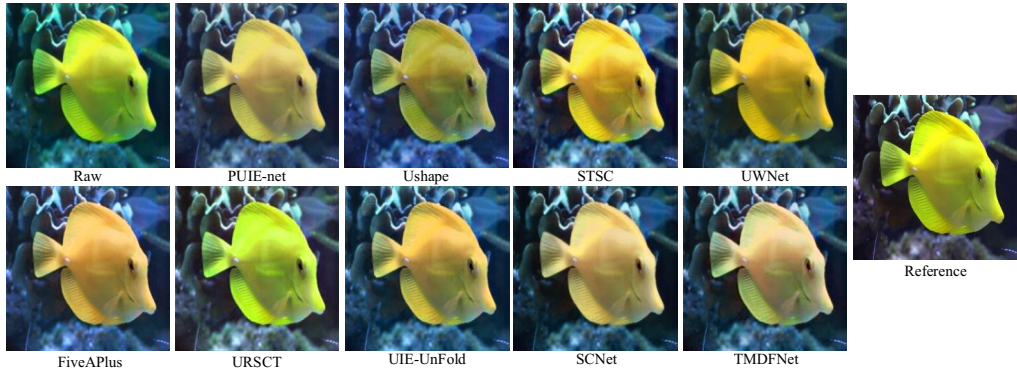


Fig. 9. The results of different methods on the EUVP.

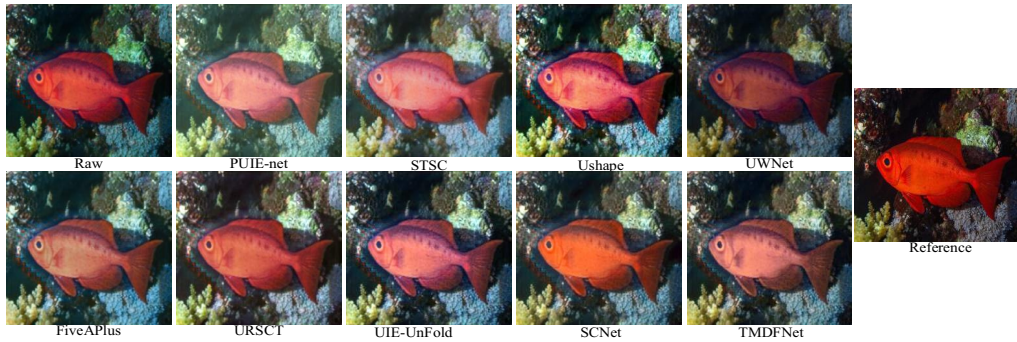


Fig. 10. The results of different methods on the UFO-120.

duced. SCNet excessively removes colors, leading to decreased contrast. In contrast, TMDFNet improves the contrast in complex environments, enhances local details, and restores relatively realistic colors.

Figure 10 shows the experimental results on the Test-U120. UWNet generates a darkened image contrast. Although URSCT and STSC can eliminate color bias, they still encounter blurring problems on fish belly details. Conversely, PUIE-net tends to remove color deviation but brings excessive image distortion. Though Ushape and UIE-UnFold can remove color deviation, it often leads to the loss of local details. SCNet completely loses local details on the belly of the fish. FiveAPlus enhances colors but produces unrealistic images. Unlike the comparison methods, TMDFNet effectively enhances contrast and details in complex scenes, resulting in images with natural colors and improved contrast.

Figure 11 shows the experimental results on Test-L900. The images of UWNet contain a lot of noise and have serious color deviation. FiveAPlus loses its original colors, resulting in image distortion. Although URSCT and UIE-UnFold restore a clear image, they do not resolve the green bias caused by light scattering. PUIE-net introduces a blue color while enhancing the image, exacerbating the color cast effect. STSC, SCNet, and

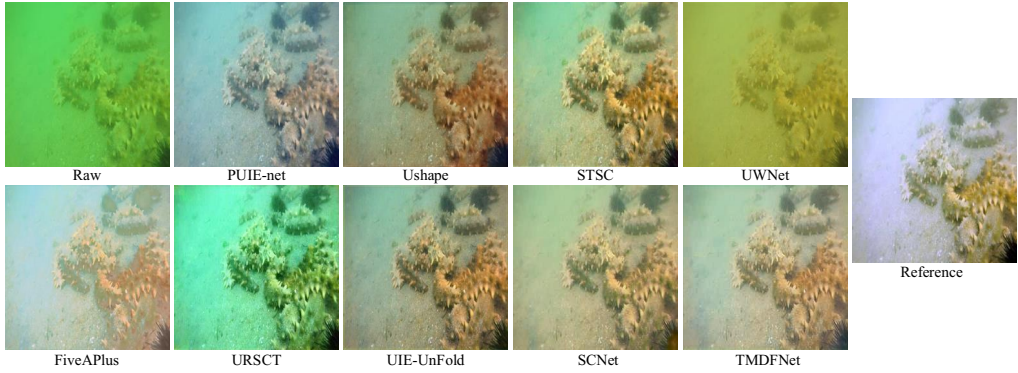


Fig. 11. The results of different methods on the LSUI.

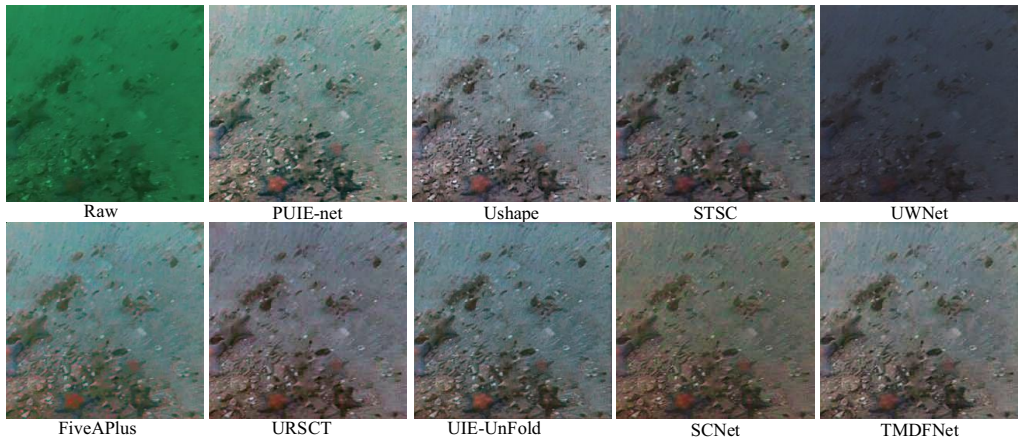


Fig. 12. The results of different methods on the RUIE.

UShape not only fail to correct the color bias in green tones, but they also exhibit excessively dark contrast. In contrast, TMDFNet models color channels and spatial regions so that recovered images have fewer color artifacts and higher fidelity that are closest to the ground truth images.

Figure 12 shows the experimental results on the Test-R100. Although UWNet removes the green hue of the image, the contrast of the image is darker. SCnet, PUIE-net, FiveAPlus and UIE-UnFold recover a green bias and false colors. STSC and URSCT remove the green bias, but the image contrast is low. Ushape can reduce the green bias to a certain extent, but the overall enhancement is not obvious. In contrast, TMDFNet considers color channels and spatial regions when processing images so that the recovered images have fewer color artifacts and higher fidelity.

4.4.2. Quantitative comparison results

The evaluation results of the reference-free metric and reference metric are shown in Table 2. Though UWNet has the highest PSNR value in Test-U512 and Test-U120, it

T a b l e 2. Performance comparisons on the EUVP, UFO, LSUI, and RUIE datasets. Bold means the best, and underlined denotes the second best. The symbols ↑ and ↓ indicate a higher score is better and a lower score is better, respectively.

Methods	U-shape	PUIE-net	STSC	UWnet	FiveAPlus	URSCT	UIE-UnFold	SCnet	TMDFNet
Test -E512	PSNR ↑	18.4893	18.6978	18.5377	22.1349	19.2454	19.4868	20.0649	<u>20.5368</u>
	SSIM ↑	0.7543	0.7879	0.7750	0.8009	0.7879	0.7828	<u>0.8047</u>	0.8095
	LPIPS ↓	0.3093	0.2943	0.2834	0.2677	0.2869	0.2856	<u>0.2760</u>	0.2800
	UIQM ↑	<u>1.5680</u>	1.4440	1.4590	1.1310	1.3630	1.4737	1.5090	1.5740
	Twice Mixing ↑	1.5391	1.4996	<u>1.5583</u>	1.2377	1.5867	<u>1.6497</u>	1.4062	1.4516
Test -U120	PSNR ↑	17.6611	17.7832	18.3644	21.7860	18.5266	19.2509	19.9484	<u>20.4391</u>
	SSIM ↑	0.5900	0.6293	0.7487	0.7644	0.6361	0.7507	<u>0.7744</u>	0.7944
	LPIPS ↓	0.6236	0.5939	0.3131	0.2937	0.6048	0.2526	0.2755	<u>0.2874</u>
	UIQM ↑	1.4580	1.4060	1.5270	1.2490	<u>1.5230</u>	1.4784	1.4470	1.5280
	Twice Mixing ↑	1.7762	1.6709	1.7587	1.4695	1.8025	1.7487	1.7702	1.8088
Test -L900	PSNR ↑	22.3069	21.3101	17.6527	16.2324	15.0281	21.6697	21.0576	<u>21.7090</u>
	SSIM ↑	0.8117	21.3101	0.4391	0.6727	0.7520	0.8219	<u>0.8464</u>	0.8528
	LPIPS ↓	0.3212	<u>0.2036</u>	0.2626	0.5227	0.4357	0.2781	0.2290	0.1936
	UIQM ↑	<u>1.6130</u>	1.5230	<u>1.6130</u>	1.6610	1.2950	1.4998	1.4990	1.6060
	Twice Mixing ↑	<u>0.9294</u>	0.6805	0.6643	0.4310	0.9550	0.9342	0.7691	0.8361
Test -R100	UIQM ↑	1.5730	1.5500	1.5740	<u>1.5980</u>	1.5530	1.5493	1.4780	1.6120
	Twice Mixing ↑	<u>0.3382</u>	0.3037	0.2734	-0.0076	0.2138	0.2932	0.1490	0.3573

is the worst from the visual effect of Fig. 9 and Fig. 10, and its reference-free metric UIQM value is also the lowest. Moreover, in Test-L900, Ushape demonstrates the highest PSNR and UIQM values in Test-L900, while the FiveAplus records the highest Twice Mixing scores in both Test-E512 and Test-L900. However, it can be observed that TMDFNet generates images that are closest to the ground truth images, as shown in Fig. 11. Besides, TMDFNet has the highest reference-free index on Test-R100, and excels in all metrics, particularly in SSIM, LPIPS, UIQM and Twice Mixing, where it consistently achieves the best or near-best scores. This demonstrates the strong capability and advantages of TMDFNet in preserving image details, perceptual quality, and underwater image processing.

Consequently, both qualitative and quantitative experiment results show that TMDFNet exhibits better comprehensive performance than the SOTA comparison methods, producing bright and saturated images that are more aligned with human visual perception. Additionally, these results highlight TMDFNet’s exceptional performance and generalization ability.

4.5. Ablation experiment

To demonstrate the effectiveness of each part, we conduct a series of ablation experiments on the UIEB dataset. The SCnet (SC) [49] is used as the baseline model. The reason is that the backbone of TMDFNet is like SC. Note that SC combines a simple U-Net with spatial and channel normalization to handle diverse water types in underwater environments. Five groups of ablation experiments are implemented, including SC, SC+TM, SC+TM+ MCEM, SC+TM+MCEM+SDFFM, and TMDFNet. As shown in Table 3, SC [49] performs the worst in terms of PSNR, SSIM, and LPIPS. With the addition of the TM, the PSNR and SSIM values of the model grow, and the LPIPS is improved. Additionally, when adding the MCEM and SDFFM, fluctuations are observed in SSIM, and both PSNR and LPIPS values show consistent improvements. Although the SSIM value is not optimal, TMDFNet gains the optimum PSNR and LPIPS. Table 3 shows that MCEM, TM, and SDFFM play crucial roles in TMDFNet. To visually demonstrate the performance of each module, two sets of ablation experiments are conducted. One set tests the texture detail recovery effect, and the other checks the color correction effect.

Table 3. Ablation on the UIEB. Bold means the best, and underlined denotes the second best. The symbols $\uparrow$ and $\downarrow$ indicate a higher score is better and a lower score is better, respectively.

Sampling methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
SC [39]	20.1793	0.8037	0.2847
SC+TM	21.8811	0.8963	0.1177
SC+TM+MCEM	21.8830	0.8921	0.1185
SC+TM+MCEM+SDFFM	<u>21.9378</u>	0.8920	<u>0.1168</u>
TMDFNet	22.0483	<u>0.8924</u>	0.1168

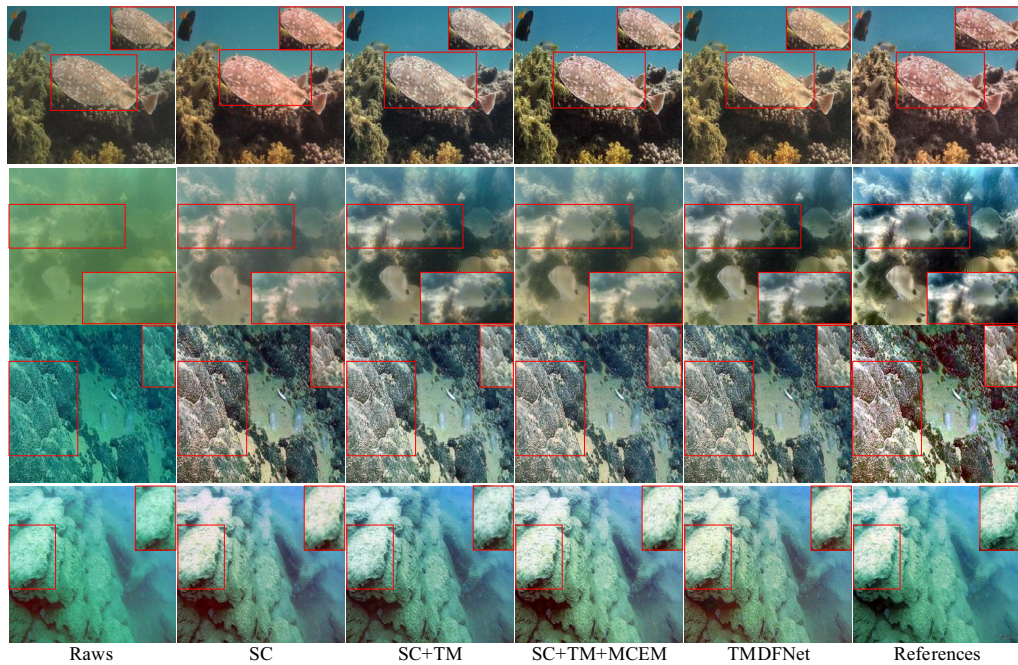


Fig. 13. Ablation experiment results of texture detail recovery effects on UIEB.

The results of the first set ablation experiment are shown in Fig. 13. Although the baseline model SC significantly reduces noise, it loses local details. Compared to the SC model, the results of SC+TM show less noise while restoring details and texture. The reason is that TM can capture complex spatial relationships in images, injecting deeper contextual information for image quality enhancement. In addition, adding MCEM expands the receptive field of the framework. Compared with SC+TM, SC+TM+MCEM pays more attention to image texture details and better processes local information, making the contrast and brightness of the image higher than SC. After adding the SDFFM module, TMDFNet contains local information and achieves the best contrast and brightness effects in the restored image. The integration of these three modules enhances the performance of TMDFNet, helping the network to recover more texture details, thereby enhancing the overall quality of degraded underwater images.

The results of the second set ablation experiment are shown in Fig. 14. The SC exhibits the worst color correction effect. Compared with SC, the results of SC+TM show significant improvement in color errors. The TM can capture the long-distance dependencies between pixels in underwater images, enhancing the understanding of the overall context of the image and thereby better alleviating color distortion caused by light scattering in the underwater environment. Compared with SC+TM, SC+TM+MCEM reduces local artifacts in the image and eliminates the green hue caused by light scattering. This enhancement is attributed to MCEM. By fusing features at different scales,

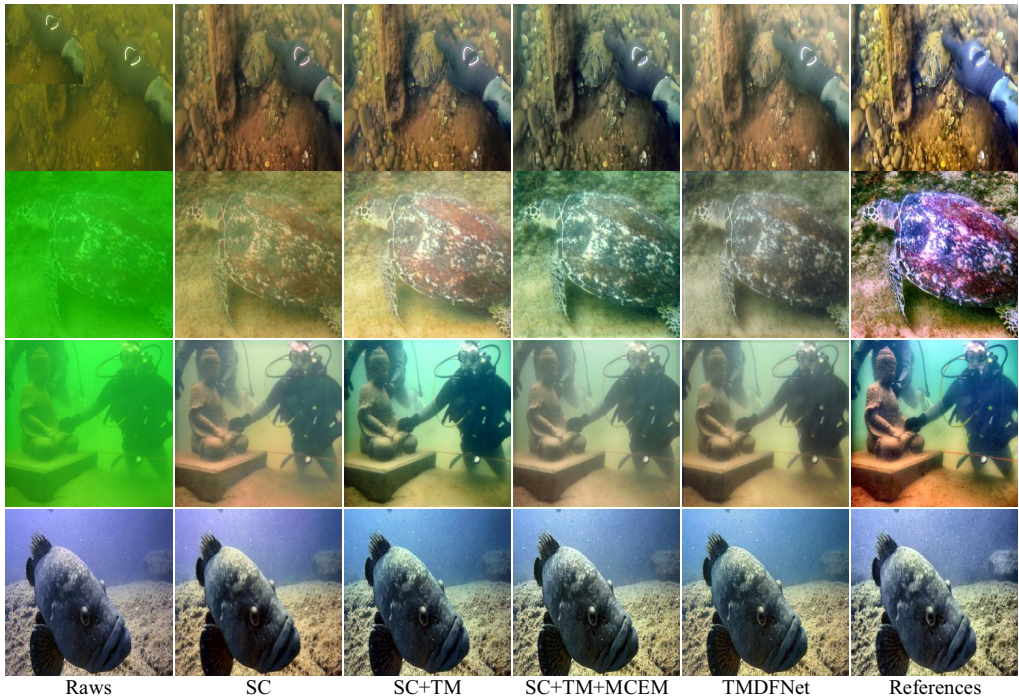


Fig. 14. Ablation experiment results on color correction effect comparison on UIEB.

detailed information of the image can be effectively retained, avoiding the loss of important details during the enhancement process, thereby improving the clarity and quality of the image. Finally, the added SDDFM calculates features in the local neighborhood of pixels, thereby improving calculation and feature fusion efficiency and achieving the best color correction effect in the restored images.

4.6. Application experiments

To highlight the superiority of the TMDFNet, we apply it to the object detection task, which utilizes the open-source dataset UTDAC2020 [55]. As shown in Fig. 15, the left image displays the original image and the object detection results of the YOLOV10 model without any preprocessing, while the right part shows the images enhanced by PUIE-Net, Ushape, STSC, UWNNet, FiveAPlus, URSCT, SCNet, and TMDFNet, with detection results using YOLOV10. The results show that the preprocessing enhancement of TMDFNet allows for the detection of targets that other SOTA algorithms cannot identify and significantly improves the detection performance of certain targets (*e.g.*, starfish and holothurian). In addition, it can be seen from Table 4 that our algorithm shows the best performance in evaluation indicators such as precision, recall, mAP50 and mAP50-95.

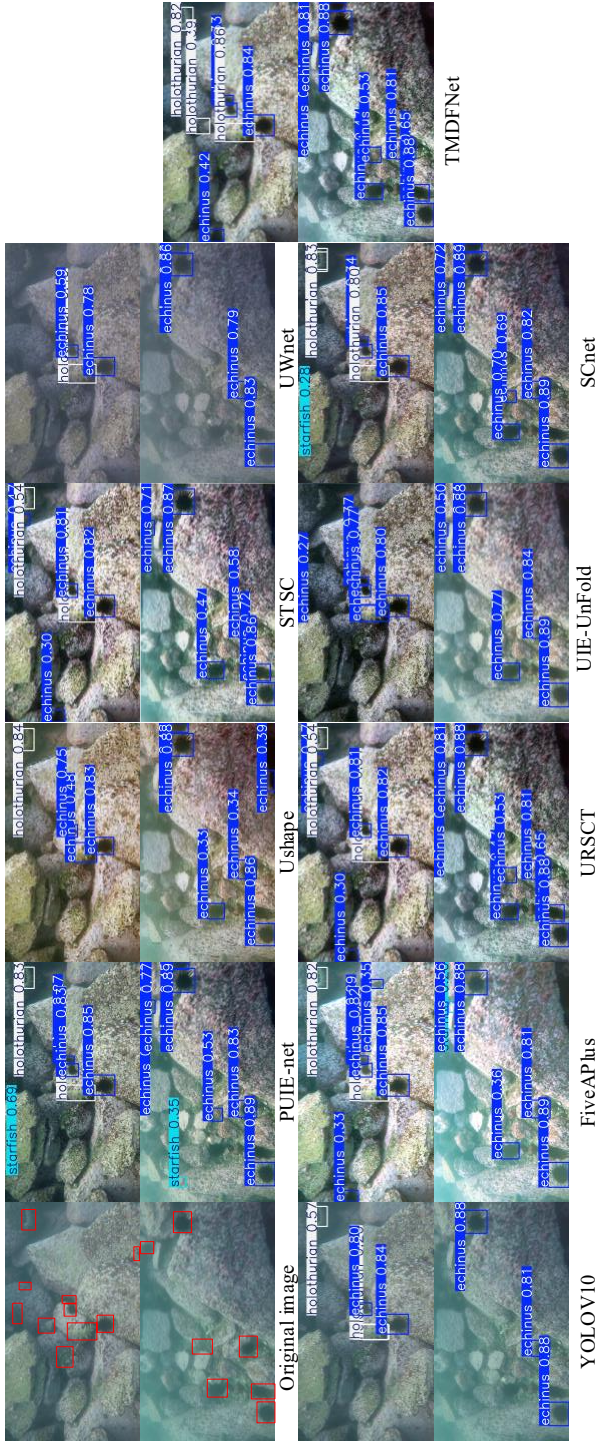


Fig. 15. Target detection application experiment on underwater dataset UTDAC2020.

T a b l e 4. Performance comparisons of target detection on the UTDAC2020 dataset. Bold means the best, and underlined denotes the second best. The symbols $\uparrow$ and $\downarrow$ indicate a higher score is better and a lower score is better, respectively.

Methods	Yolov10	U-shape	PUIE-net	STSC	UWnet	FiveAPlus	URSCT	UIE-UnFold	SCnet	TMDFNet
Precision $\uparrow$	0.744	0.725	0.756	0.758	0.714	0.749	0.742	0.727	<u>0.767</u>	0.795
Recall $\uparrow$	0.609	0.543	<u>0.690</u>	0.628	0.502	0.593	0.653	0.610	0.674	0.733
mAP50 $\uparrow$	0.670	0.612	<u>0.745</u>	0.720	0.575	0.661	0.729	0.664	0.737	0.795
Map50-95 $\uparrow$	0.366	0.322	0.411	0.435	0.308	0.351	<u>0.436</u>	0.348	0.403	0.458

5. Conclusion

We have proposed a UIE method with a Transformer-based multi-scale spatial domain feature fusion network (TMDFNet). TMDFNet integrates both high-level and low-level information to enhance image details by establishing connections through feature fusion. An MCEM is designed to enhance features at different scales, facilitating the learning of texture details at various hierarchical levels through context links and expanding the depth and width of the network. An SDFFM is developed to enhance global information in the image from the Fourier domain perspective. A TM is created to enhance the overall understanding of information in the network by capturing complex relationships in the image. Additionally, an attention mechanism is introduced to perceive color changes in underwater images across different channels and spatial regions. With the integration of these modules, combined with the perceptual similarity loss function, the contrast and saturation of the restored images are further enhanced. In the experiments, three evaluation metrics are employed for qualitative and quantitative comparisons, and the results confirm the outstanding performance of the TMDFNet. Moreover, ablation experiments validate the effectiveness of each module in TMDFNet and indicate their contributions to the network. In the future, we will prioritize improving computational efficiency, particularly by optimizing training for resource-constrained environments. Although TMDFNet makes significant progress in UIE, it still faces challenges under extreme conditions such as severe light scattering, color distortion, and suspended particles, coupled with high computational demands. Therefore, we plan to simplify the model and bolster its robustness in these complex scenarios.

Funding

This work was supported by the Guangdong Basic and Applied Basic Research Foundation Project (2025A1515140054, 2025A1515010022), National Natural Science Foundation of China (62573140, 62673139, 62173098), Shenzhen Science and Technology Program (JCYJ20250604174522029), Scientific Research Innovation Capability Support Project for Young Faculty (SRICSPYF-BS2025018), and Guangdong Provincial Key Laboratory of Cyber-Physical System (2020B1212060069).

Disclosures

The authors declare no conflicts of interest.

References

- [1] SAHU P., GUPTA N., SHARMA N., *A survey on underwater image enhancement techniques*, International Journal of Computer Applications **87**(13), 2014: 19–23. <https://doi.org/10.5120/15268-3743>
- [2] DE SIQUEIRA NETO A.C., PORSANI J.L., RANGEL R.C., DE SOUZA L.A.P., NAVARRO A.G., DE LIMA L.G., STANGARI M.C., *Geophysical surveys at Formoso underwater archaeological stilt village in the eastern Amazon region, Brazil*, Journal of Archaeological Science: Reports **60**, 2024: 104821. <https://doi.org/10.1016/j.jasrep.2024.104821>
- [3] YOUSRI D., ABD ELAZIZ M., OLIVA D., ABRAHAM A., ALOTAIBI M.A., HOSSAIN M.A., *Fractional-order comprehensive learning marine predators algorithm for global optimization and feature selection*, Knowledge-Based Systems, **235**, 2022: 107603. <https://doi.org/10.1016/j.knosys.2021.107603>
- [4] WU X., CHEN H., CHEN C., ZHONG M., XIE S., GUO Y., FUJITA H., *The autonomous navigation and obstacle avoidance for USVs with ANOA deep reinforcement learning method*, Knowledge-Based Systems, **196**, 2020: 105201. <https://doi.org/10.1016/j.knosys.2019.105201>
- [5] SOARES M.S., SINGH R., KUMAR S., JHA R., NEDOMA J., MARTINEK R., MARQUES C., *The role of smart optical biosensors and devices on predictive analytics for the future of aquaculture systems*, Optics & Laser Technology, **177**, 2024: 111049. <https://doi.org/10.1016/j.optlastec.2024.111049>
- [6] XIANG K., DONG X., WANG W., DONG X., *Edge-guided oceanic scene element detection*, Knowledge-Based Systems, **284**, 2024: 111322. <https://doi.org/10.1016/j.knosys.2023.111322>
- [7] MEI J., WU L., CHEN E., XIAO W., ZHONG L., GUO J., LI W., *A novel structural damage detection method using a hybrid IDE–BP model*, Knowledge-Based Systems, **273**, 2023: 110606. <https://doi.org/10.1016/j.knosys.2023.110606>
- [8] DUNTLEY S.Q., *Light in the sea*, Journal of the Optical Society of America, **53**(2), 1963: 214–233. <https://doi.org/10.1364/JOSA.53.000214>
- [9] ZHAO W., RONG S., FENG C., HE B., *PSNet: A non-uniform illumination correction method for underwater images based pseudo-siamese network*, Knowledge-Based Systems, **309**, 2025: 112780. <https://doi.org/10.1016/j.knosys.2024.112780>
- [10] LI Y., LI D., GAO Z., WANG S., JIAO Q., BIAN L., *Underwater image enhancement utilizing adaptive color correction and model conversion for dehazing*, Optics & Laser Technology, **169**, 2024: 110039. <https://doi.org/10.1016/j.optlastec.2023.110039>
- [11] BERMAN D., LEVY D., AVIDAN S., TREIBITZ T., *Underwater single image color restoration using haze-lines and a new quantitative dataset*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **43**(8), 2021: 2822–2837. <https://doi.org/10.1109/TPAMI.2020.2977624>
- [12] AKKAYNAK D., TREIBITZ T., SHLESINGER T., LOYA Y., TAMIR R., ILUZ D., *What is the space of attenuation coefficients in underwater computer vision?*, [In] *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017: 568–577. <https://doi.org/10.1109/CVPR.2017.68>
- [13] WANG R., ZHANG Y., ZHANG J., *An efficient swin transformer-based method for underwater image enhancement*, Multimedia Tools and Applications, **82**, 2023: 18691–18708. <https://doi.org/10.1007/s11042-022-14228-6>
- [14] LU J., LI N., ZHANG S., YU Z., ZHENG H., ZHENG B., *Multi-scale adversarial network for underwater image restoration*, Optics & Laser Technology, **110**, 2019: 105–113. <https://doi.org/10.1016/j.optlastec.2018.05.048>
- [15] HUO F., LI B., ZHU X., *Efficient wavelet boost learning-based multi-stage progressive refinement network for underwater image enhancement*, [In] *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, 2021: 1944–1952. <https://doi.org/10.1109/ICCVW54120.2021.00221>
- [16] SINGH N., BHAT A., *Underwater image enhancement using convolutional block attention module*, [In] *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, IEEE, Mathura, India, 2023: pp. 1–5. <https://doi.org/10.1109/ISCON57294.2023.10111974>

- [17] BADRAN M., TORKI M., *DAUT: Underwater image enhancement using depth aware U-shape transformer*, [In] *2023 IEEE International Conference on Image Processing (ICIP)*, Kuala Lumpur, Malaysia, 2023: 1830–1834. <https://doi.org/10.1109/ICIP49359.2023.10222895>
- [18] LIU Y., XU H., ZHANG B., SUN K., YANG J., LI B., LI C., QUAN X., *Model-based underwater image simulation and learning-based underwater image enhancement method*, *Information*, **13**(4), 2022: 187. <https://doi.org/10.3390/info13040187>
- [19] LIU Y., XU H., SHANG D., LI C., QUAN X., *An underwater image enhancement method for different illumination conditions based on color tone correction and fusion-based descattering*, *Sensors*, **19**(24), 2019: 5567. <https://doi.org/10.3390/s19245567>
- [20] PORTO MARQUES T., BRANZAN ALBU A., HOEBERECHTS M., *A contrast-guided approach for the enhancement of low-lighting underwater images*, *Journal of Imaging*, **5**(10), 2019: 79. <https://doi.org/10.3390/jimaging5100079>
- [21] LIU Y., YAN Z., TAN J., LI Y., *Multi-purpose oriented single nighttime image haze removal based on unified variational retinex model*, *IEEE Transactions on Circuits and Systems for Video Technology*, **33**(4), 2023: 1643–1657. <https://doi.org/10.1109/TCSVT.2022.3214430>
- [22] WU C., YU S., HU Q., XU J., ZHANG L., *Underwater image enhancement via dehazing and color restoration*, 2024, arXiv:2409.09779v1. <https://doi.org/10.48550/arXiv.2409.09779>
- [23] TANG P., LI L., XUE Y., LV M., JIA Z., MA H., *Real-world underwater image enhancement based on attention U-net*, *Journal of Marine Science and Engineering*, **11**(3), 2023: 662. <https://doi.org/10.3390/jmse11030662>
- [24] HUANG Z., LI J., HUA Z., FAN L., *Underwater image enhancement via adaptive group attention-based multiscale cascade transformer*, *IEEE Transactions on Instrumentation and Measurement*, **71**, 2022: 1–18. <https://doi.org/10.1109/TIM.2022.3189630>
- [25] QI Q., LI K., ZHENG H., GAO X., HOU G., SUN K., *SGUIE-Net: Semantic attention guided underwater image enhancement with multi-scale perception*, *IEEE Transactions on Image Processing*, **31**, 2022: 6816–6830. <https://doi.org/10.1109/TIP.2022.3216208>
- [26] TIAN Y., XU Y., ZHOU J., *Underwater image enhancement method based on feature fusion neural network*, *IEEE Access*, **10**, 2022: 107536–107548. <https://doi.org/10.1109/ACCESS.2022.3210941>
- [27] MI Z., LIANG Z., WANG Y., FU X., CHEN Z., *Multi-scale gradient domain underwater image enhancement*, [In] *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, IEEE, Kobe, Japan, 2018: 1–5. <https://doi.org/10.1109/OCEANSKOB.2018.8559180>
- [28] FAN X., LU L., SHI P., ZHANG X., *Underwater image enhancement algorithm combining color correction and multi-scale fusion*, [In] *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, IEEE, Chiang Mai, Thailand, 2021: 140–143. <https://doi.org/10.1109/ECTI-CON51831.2021.9454733>
- [29] CHEN R., CAI Z., CAO W., *MFFN: An underwater sensing scene image enhancement method based on multiscale feature fusion network*, *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 2022: 4205612. <https://doi.org/10.1109/TGRS.2021.3134762>
- [30] UEKI Y., IKEHARA M., *Underwater image enhancement with multi-scale residual attention network*, [In] *2021 International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, Munich, Germany, 2021: 1–5. <https://doi.org/10.1109/VCIP53242.2021.9675342>
- [31] QIAO N., DONG L., SUN C., *Adaptive deep learning network with multi-scale and multi-dimensional features for underwater image enhancement*, *IEEE Transactions on Broadcasting*, **69**(2), 2023: 482–494. <https://doi.org/10.1109/TBC.2022.3227424>
- [32] YANG G., KANG G., LEE J., CHO Y., *Joint-ID: Transformer-based joint image enhancement and depth estimation for underwater environments*, *IEEE Sensors Journal*, **24**(3), 2024: 3113–3122. <https://doi.org/10.1109/JSEN.2023.3338730>
- [33] LIU J., LUO X., WANG D., LI F., LI J., LAN R., *MobileVitV2-based fusion of vision transformers and convolutional neural networks for underwater image enhancement*, [In] *2023 13th International Conference on Information Science and Technology (ICIST)*, IEEE, Cairo, Egypt, 2023: 195–204. <https://doi.org/10.1109/ICIST59754.2023.10367056>

- [34] REN T., XU H., JIANG G., YU M., ZHANG X., WANG B., LUO T., *Reinforced swin-convs transformer for simultaneous underwater sensing scene image enhancement and super-resolution*, IEEE Transactions on Geoscience and Remote Sensing, **60**, 2022: 4209616. <https://doi.org/10.1109/TGRS.2022.3205061>
- [35] WU J., LUO T., HE Z., SONG Y., XU H., LI L., *CEWformer: A transformer-based collaborative network for simultaneous underwater image enhancement and watermarking*, IEEE Journal of Oceanic Engineering, **49**(1), 2024: 30–47. <https://doi.org/10.1109/JOE.2023.3310079>
- [36] JAFFE J.S., *Computer modeling and the design of optimal underwater imaging systems*, IEEE Journal of Oceanic Engineering, **15**(2), 1990: 101–111. <https://doi.org/10.1109/48.50695>
- [37] ZHOU J., LI B., ZHANG D., YUAN J., ZHANG W., CAI Z., SHI J., *UGIF-Net: An efficient fully guided information flow network for underwater image enhancement*, IEEE Transactions on Geoscience and Remote Sensing, **61**, 2023: 4206117. <https://doi.org/10.1109/TGRS.2023.3293912>
- [38] LI B., WU F., WEINBERGER K.Q., BELONGIE S., *Positional normalization*, 2019, arXiv:1907.04312. (accessed May 14, 2024). <https://doi.org/10.48550/arXiv.1907.04312>
- [39] FU Z., LIN X., WANG W., HUANG Y., DING X., *Underwater image enhancement via learning water type desensitized representations*, [In] *ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Singapore, Singapore, 2022: 2764–2768. <https://doi.org/10.1109/ICASSP43922.2022.9747758>
- [40] HU J., SHEN L., ALBANIE S., SUN G., WU E., *Squeeze-and-excitation networks*, 2019, arXiv:1709.01507 (accessed January 2, 2024). <https://doi.org/10.48550/arXiv.1709.01507>
- [41] JIANG J., YE T., BAI J., CHEN S., CHAI W., JUN S., LIU Y., CHEN E., *Five A⁺ network: You only need 9K parameters for underwater image enhancement*, 2023, arXiv:2305.08824 (accessed December 22, 2023). <https://doi.org/10.48550/arXiv.2305.08824>
- [42] GAO Y., ZHOU M., METAXAS D., *UTNet: A hybrid transformer architecture for medical image segmentation*, 2021, arXiv:2107.00781 (accessed December 20, 2023). <https://doi.org/10.48550/arXiv.2107.00781>
- [43] LI C., GUO C., REN W., CONG R., HOU J., KWONG S., TAO D., *An underwater image enhancement benchmark dataset and beyond*, IEEE Transactions on Image Processing, **29**, 2020: 4376–4389. <https://doi.org/10.1109/TIP.2019.2955241>
- [44] ISLAM M.J., XIA Y., SATTAR J., *Fast underwater image enhancement for improved visual perception*, IEEE Robotics and Automation Letters, **5**(2), 2020: 3227–3234. <https://doi.org/10.1109/LRA.2020.2974710>
- [45] ISLAM M.J., LUO P., SATTAR J., *Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception*, [In] *Robotics: Science and Systems 2020*, Corvallis, Oregon, USA, July 12–16, 2020. <https://doi.org/10.15607/RSS.2020.XVI.018>
- [46] LIU R., FAN X., ZHU M., HOU M., LUO Z., *Real-world underwater enhancement: Challenges, benchmarks, and solutions*, 2019, arXiv:1901.05320 (accessed May 6, 2024). <https://doi.org/10.48550/arXiv.1901.05320>
- [47] HORÉ A., ZIOU D., *Image quality metrics: PSNR vs. SSIM*, [In] *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010: 2366–2369. <https://doi.org/10.1109/ICPR.2010.579>
- [48] ZHANG R., ISOLA P., EFROS A.A., SHECHTMAN E., WANG O., *The unreasonable effectiveness of deep features as a perceptual metric*, [In] *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 586–595. <https://doi.org/10.1109/CVPR.2018.00068>
- [49] PANETTA K., GAO C., AGAIAN S., *Human-visual-system-inspired underwater image quality measures*, IEEE Journal of Oceanic Engineering, **41**(3), 2016: 541–551. <https://doi.org/10.1109/JOE.2015.2469915>
- [50] FU Z., FU X., HUANG Y., DING X., *Twice mixing: A rank learning based quality assessment approach for underwater image enhancement*, Signal Processing: Image Communication **102**, 2022: 116622. <https://doi.org/10.1016/j.image.2021.116622>

- [51] FU Z., WANG W., HUANG Y., DING X., MA K.-K., *Uncertainty inspired underwater image enhancement*, 2022, arXiv:2207.09689 (accessed December 22, 2023). <https://doi.org/10.48550/arXiv.2207.09689>
- [52] WANG D., MA L., LIU R., FAN X., *Semantic-aware texture-structure feature collaboration for underwater image enhancement*, [In] *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, Philadelphia, PA, USA, 2022: 4592–4598. <https://doi.org/10.1109/ICRA46639.2022.9812457>
- [53] NAIK A., SWARNAKAR A., MITTAL K., *Shallow-UWnet : Compressed model for underwater image enhancement*, 2021, arXiv:2101.02073 (accessed December 22, 2023). <https://doi.org/10.48550/arXiv.2101.02073>
- [54] LEI Y., YU J., DONG Y., GONG C., ZHOU Z., PUN C.-M., *UIE-UnFold: Deep unfolding network with color priors and vision transformer for underwater image enhancement*, [In] *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, San Diego, CA, USA, 2024: 1–10. <https://doi.org/10.1109/DSAA61799.2024.10722842>
- [55] SONG P., LI P., DAI L., WANG T., CHEN Z., *Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection*, *Neurocomputing*, **530**, 2023: 150–164. <https://doi.org/10.1016/j.neucom.2023.01.088>

*Received October 23, 2025
in revised form February 19, 2026*