MAAU-UIE: Multiple Attention Aggregation U-Net for Underwater Image Enhancement

Junsheng Chang, Qin Shi, Yijun Zhang, Zongtang Hu
China Mobile (Suzhou) Software Technology Co., Ltd.
Suzhou, Jiangsu, China

Xulun Ye
Faculty Of Electrical Engineering and Computer Science, Ningbo University
Ningbo, Zhejiang, China

Abstract

To address issues such as color distortion, blurriness, and low contrast in underwater images, a Multiple Attention Aggregation U-Net for Underwater Image Enhancement (MAAU-UIE) is proposed. The network is constructed using an encoder-decoder structure, with a multiple attention block designed to enhance the ability to extract features from low-quality underwater images. First, axial rectangular window attention and shifted axial rectangular window attention are alternately applied to learn local context and establish global dependencies, respectively. Additionally, channel convolution and spatial convolution are incorporated into the process of window attention calculation to further supplement local information. A channel enhancement module is then added to improve the modeling capability in the channel dimension. Finally, gradient loss and multi-scale structural similarity loss are used to enhance the network's ability to extract edge detail information and multi-scale structural features. Ablation experiments demonstrate the significant role of each proposed module plays in improving network performance. Quantitative experiments show that this method surpasses existing methods on various objective metrics. The peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) on the benchmark dataset UIEB test set reach 24.467 and 0.920, respectively. The underwater image quality measurement (UIQM) and underwater color image quality evaluation (UCIQE) on the UIEB challenge set and in three color-bias environments of UCCS outperform cutting-edge methods. Qualitative experiments indicate a clear advantage in subjective visual effects, effectively restoring underwater images with natural colors and clear texture structures.

***Keywords:** underwater image enhancement, attention aggregation, rectangular window attention, channel feature enhancement, multi-scale*

1. Introduction

With the continuous advancements in ocean exploration and underwater activities, the acquisition of underwater images has become increasingly important in fields such as marine science, archaeology, ocean engineering, and underwater robot navigation [30, 25]. However, due to the complexity of the underwater environment, underwater images are often affected by factors such as light scattering and absorption [20], which lead to poor image quality. These issues mainly manifest as color distortion, reduced contrast, blurring, and increased noise [1, 2], significantly lowering the usability and visual quality of underwater images. As a result, underwater image enhancement technologies have emerged, aiming to improve the visual quality of underwater images through algorithms and technical approaches, making them more suitable for human perception and computer vision applications.

Traditional methods [1, 2, 34] rely on statistical information and assumptions about the image. These methods attempt to improve visual quality by correcting color distortions and enhancing contrast through manually designed features. However, these methods are unable to adapt to dynamic scenes and perform poorly in terms of image quality restoration.

Convolutional neural networks (CNNs) perform excellently in feature extraction for images and are therefore widely used in image restoration and related fields. CNN-based methods [5, 12, 13] can automatically extract image features and achieve end-to-end image representation. However, the receptive field and fixed convolutional kernels of CNN-based methods limit their development in the field of image restoration. Due to the limitations of the receptive field, CNN-based methods cannot capture relationships between pixels over a larger range. Furthermore, the fixed convolutional kernels prevent CNN-based methods from adapting to the diversity of underwater images.

The transformer model, based on self-attention, was initially applied in natural language processing (NLP) [10]. In recent years, due to its outstanding performance in visual

tasks, many transformer-based methods [17, 19] have been applied to image restoration. However, due to its quadratic computational complexity, it faces challenges when dealing with high-resolution images, as the computational load becomes enormous and needs to be optimized.

The main contributions are summarized as follows:

1) The network is built upon the U-Net architecture, which leverages hierarchical up-sampling and down-sampling operations to effectively extract information from feature maps of different scales. This enhances the model's ability to handle various complex underwater environments.

2) Standard U-Net networks, relying purely on convolutional operations, have limited ability to capture global information. To address this, the base module of the U-Net is designed using a window attention mechanism. To mitigate the high computational cost of standard Transformer models, the feature map is divided into a series of rectangular windows, and attention is calculated within each window to effectively extract information in both horizontal and vertical directions. Further, by shifting the axial rectangular window partitioning, connections between windows are strengthened.

3) To further enhance the feature extraction capabilities, convolution operations are integrated into the window attention computation to capture richer local detail information. Additionally, a channel enhancement module is added, which assigns different weight coefficients to each channel, improving modeling capacity.

4) In addition to pixel-based reconstruction loss, gradient loss and multi-scale structural similarity loss are incorporated to enhance the robustness of the model in recovering texture details.

5) Extensive experiments conducted on the UIEB benchmark dataset [14] show that our method significantly improves performance across various objective metrics and demonstrates superior subjective visual results. Furthermore, validation on the UCCS dataset [16] confirms that the proposed method achieves superior performance in underwater environments with three different color casts.

2. Related Work

In recent years, significant progress has been made in the field of underwater image enhancement. Traditional methods [2, 34]continue to be optimized and applied, but more attention is being given to deep learning-based enhancement methods, which have gradually become a research hotspot. For instance, Islam et al. [9] designed a multi-scale convolutional neural network that improves the contrast and clarity of underwater images through multi-layer feature extraction and fusion.

Moreover, Generative Adversarial Networks (GANs) have demonstrated great potential in underwater image enhancement. Ye et al. [29] utilized GANs to achieve efficient image denoising and dehazing, significantly enhancing the visual quality of underwater images. Zhang et al. [31] proposed an enhanced GAN architecture that incorporates a self-attention mechanism in both the generator and discriminator, further improving color balance and detail representation. Cong et al. [4] introduced a GAN-based physical model for underwater image enhancement, which performed well in terms of visual aesthetics. However, GAN-based methods rely on large amounts of training data, and the training process may encounter mode collapse issues, leading to instability in the quality of the generated images.

In recent years, attention mechanisms and Transformer structures have also been introduced into the field of underwater image enhancement. Huang et al. [8] introduced a Transformer-based method for underwater image enhancement, which effectively integrates global and local information in the image through self-attention mechanisms, enabling more precise correction of color and contrast issues. Ren et al. [19, 21] employed Transformers and self-attention mechanisms for underwater image enhancement, which can capture a broader range of pixel features from the input images, although these methods tend to have higher computational demands. The underwater image enhancement method based on Mamba[6] is another promising research direction. Lin et al. [15] proposed PixMamba, applying Mamba to underwater image enhancement. Mamba offers efficient sequence modeling capabilities, its primary limitation lies in the increased difficulty of model tuning due to its complexity. Furthermore, its application in the image domain remains in the exploratory phase, requiring more experimental evidence to verify its effectiveness in underwater image enhancement.

Although significant progress has been made, current underwater image enhancement methods still face several challenges. First, the diversity and complexity of underwater environments lead to inconsistent performance of existing models across different settings, lacking generalizability. Second, deep learning-based methods require substantial computational resources and depend on large amounts of annotated data, which poses challenges for real-world applications. Additionally, most existing research focuses on a single image enhancement task, and there is still no unified framework that can simultaneously handle multiple tasks such as color correction, contrast enhancement, denoising, and dehazing. Therefore, further research is necessary to develop more general, efficient, and resource-light underwater image enhancement methods.

To address the above challenges, This paper proposes a Multiple Attention Aggregation U-Net [22] for Underwater Image Enhancement (MAAU-UIE) based on multiple attention mechanisms combined with effective loss functions, which achieves promising performance in underwater image enhancement.

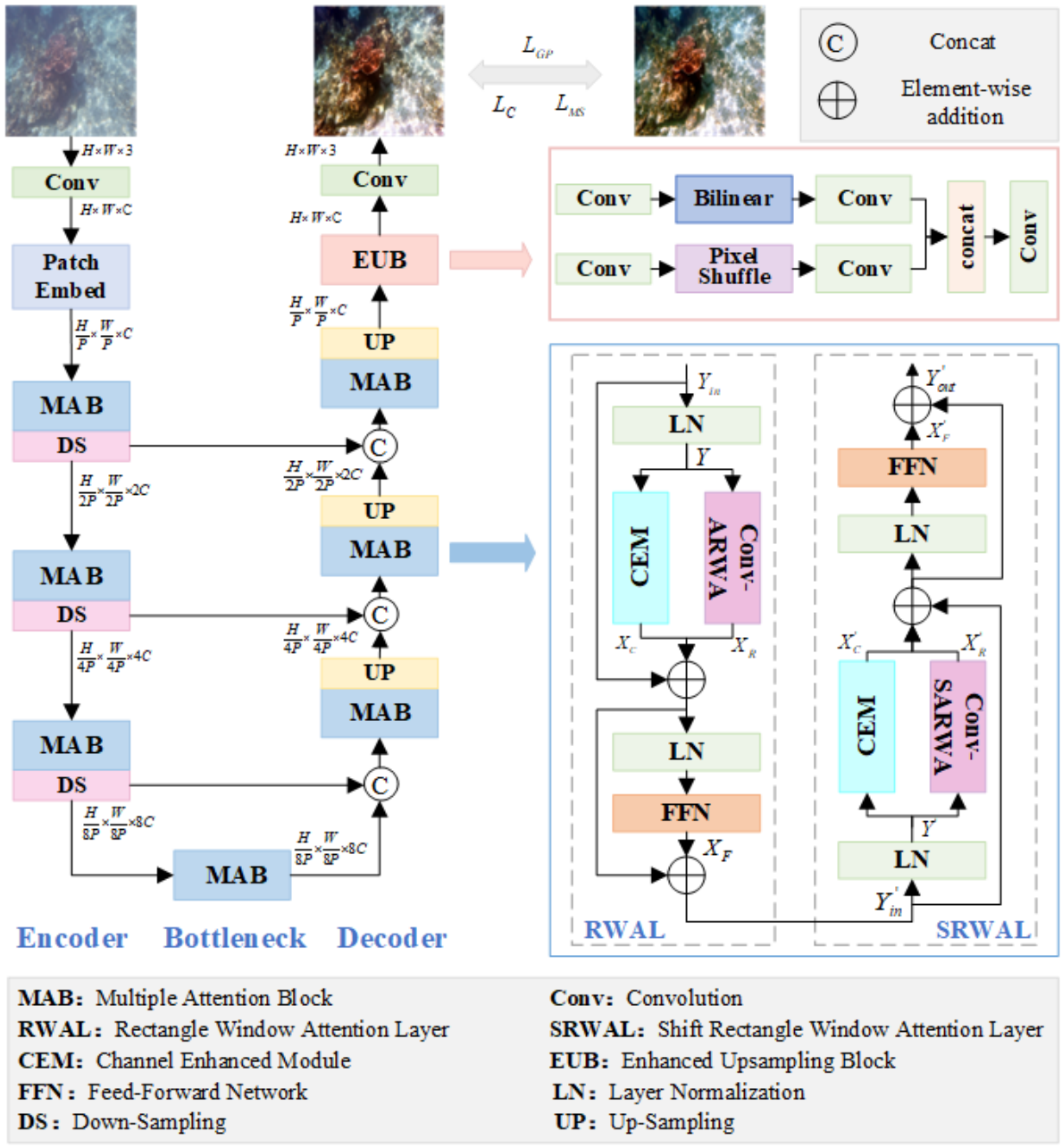

Figure 1. The network structure of MAAU-UIE

3. Methodology

3.1. Network Architecture

This paper presents a Multiple Attention Aggregation U-Net network (MAAU-UIE) for underwater image enhancement. The network structure, as shown in Fig. 1, primarily consists of an encoder, a bottleneck layer, and a decoder. A novel Multiple Attention Block (MAB) is designed as the foundational module of the network. The original underwater image is first resized to a predefined dimension, denoted as $I_{in} \in R^{H\times W\times 3}$. The image is then passed through a 3×3 convolutional layer to map the number of channels to C. After that, a PatchEmbed layer is applied to divide the image into a series of patches of size $P \times P$, generating shallow encoded features, denoted as $X_0 \in R^{\frac{H}{P}\times\frac{W}{P}\times C}$.

The encoder consists of three Multiple Attention Blocks (MABs), each followed by a PatchMerging module used as a down-sampling (DS) operation. The PatchMerging

reduces the width and height of the feature maps by half while doubling the number of channels. Taking the input X_0, the outputs of the three DS modules in the encoder are denoted as $E_1 \in R^{\frac{H}{2P} \times \frac{W}{2P} \times 2C}$, $E_2 \in R^{\frac{H}{4P} \times \frac{W}{4P} \times 4C}$ and $E_3 \in R^{\frac{H}{8P} \times \frac{W}{8P} \times 8C}$ respectively. The bottleneck layer includes one MAB that further enhances the features, producing the output $E_3^{'} \in R^{\frac{H}{8P} \times \frac{W}{8P} \times 8C}$.

The decoder has a symmetric structure to the encoder, also comprising three MABs. Each MAB in the decoder is followed by an up-sampling (UP) module, which uses bicubic interpolation to double the width and height of the feature maps while halving the number of channels. The outputs of the three UP modules in the decoder are denoted as $D_1 \in R^{\frac{H}{4P} \times \frac{W}{4P} \times 4C}$, $D_2 \in R^{\frac{H}{2P} \times \frac{W}{2P} \times 2C}$ and $D_3 \in R^{\frac{H}{P} \times \frac{W}{P} \times C}$ respectively. Simultaneously, the feature maps from the encoder and decoder at corresponding positions are concatenated along the channel dimension. A fully connected (Linear) layer is then applied to reduce the number of channels by half, improving the information flow across the network.

To retain more image details and enhance the visual quality of the reconstructed image, an Enhanced Upsampling Block (EUB) is used after the decoder. This block employs parallel sub-pixel convolution (PixelShuffle) and bilinear interpolation to double the width and height of the feature map, followed by channel concatenation. The resulting feature map is then processed through a 1×1 convolutional layer to extract spatial features and restore the number of channels to C, resulting in output features $I_F \in R^{H \times W \times C}$.At the end of the network, a 3×3 convolutional layer is applied to map the channel dimension of I_F to 3, producing the enhanced underwater image, denoted as $I_{out} \in R^{H \times W \times C}$.

3.2. Multiple Attention Block

As shown in the blue box in Fig. 1, the Multiple Attention Block (MAB) consists of the Rectangle Window Attention Layer (RWAL) and the Shift Rectangle Window Attention Layer (SRWAL). These two window attention layers alternate to learn local features and extract global contextual information. For simplicity, let the input features of MAB be denoted as $Y_{in} \in R^{H \times W \times C}$, which are parallelly input into the Convolution-Axial Rectangle Window Attention (Conv-ARWA) and the Channel Enhanced Module (CEM).

The Conv-ARWA focuses on computing correlations within rectangular windows, extracting information in both horizontal and vertical directions, while utilizing convolution layers to enhance the learning of local features. The CEM establishes correlations between feature channels, and its output is combined with the results from Conv-ARWA through a weighted summation, controlling the weights of spatial and channel features during fusion. The overall process of RWAL is described by Equations (1) to (4):

$$X_R = Conv - ARWA(LN(Y_{in}) \tag{1}$$

$$X_C = CEM(LN(Y_{in})) \tag{2}$$

$$X_F = FFN(LN(\alpha \cdot X_C + X_R + Y_{in})) \tag{3}$$

$$Y_{in}^{'} = \alpha \cdot X_C + X_R + X_F \tag{4}$$

where LN represents Layer Normalization, FFN represents the Feed-Forward Network, and α represents the channel feature weights. The Shift Rectangle Window Attention Layer (SRWAL) replaces the Conv-ARWA in RWAL with Convolution-Shift Axial Rectangle Window Attention (Conv-SARWA), while the remaining steps are identical to RWAL.

3.3. Rectangle Window Partitioning

In the standard Transformer model [24], self-attention is computed across all pixels in the entire feature map, which results in high computational costs and limited capability in extracting local features. To ensure effective feature extraction while reducing computational load, the feature map is divided into a series of non-overlapping rectangular windows as described in [3]. Let the width and height of each rectangular window be denoted as M_w and M_h, respectively. Based on the relationship between M_w and M_h, the rectangular windows are categorized into two types. If $M_w > M_h$, the window is defined as a horizontal window, and if $M_w < M_h$, it is defined as a vertical window.

As illustrated in Fig. 2, Rectangle Window Attention (RWA) divides the feature map $Y \in R^{H \times W \times C}$ along the channel dimension into two parts, $Y_1 \in R^{H \times W \times \frac{C}{2}}$ and $Y_2 \in R^{H \times W \times \frac{C}{2}}$, and further partitions Y_1 and Y_2 into a series of horizontal and vertical windows, respectively. Let Y_1 and Y_2 be considered as consisting of $1 \times 1 \times \frac{c}{2}$ tokens, each with a dimension of $H \times W$; hence, each window can be viewed as containing $M_w \times M_h$ tokens. In Fig. 2, the green boxes represent tokens, and the blue rectangular boxes represent the partitioned windows used for subsequent attention calculation. Compared to square windows [17], rectangular windows capture more information along the horizontal and vertical directions for each pixel, which helps enhance the model's feature extraction capabilities.

Considering that Rectangle Window Attention (RWA) focuses on information interaction within each window and lacks connections between different windows, Shift Rectangle Window Attention (SRWA) is added to further expand the receptive field. As shown in Fig. 2, a cyclic shifting operation is applied: the windows partitioned in RWA are shifted downward by $\frac{M_h}{2}$ pixels and rightward by $\frac{M_w}{2}$ pixels, resulting in newly partitioned windows. Finally, the results from horizontal SRWA and vertical SRWA are concatenated along the channel dimension to obtain the final output of the window attention mechanism.

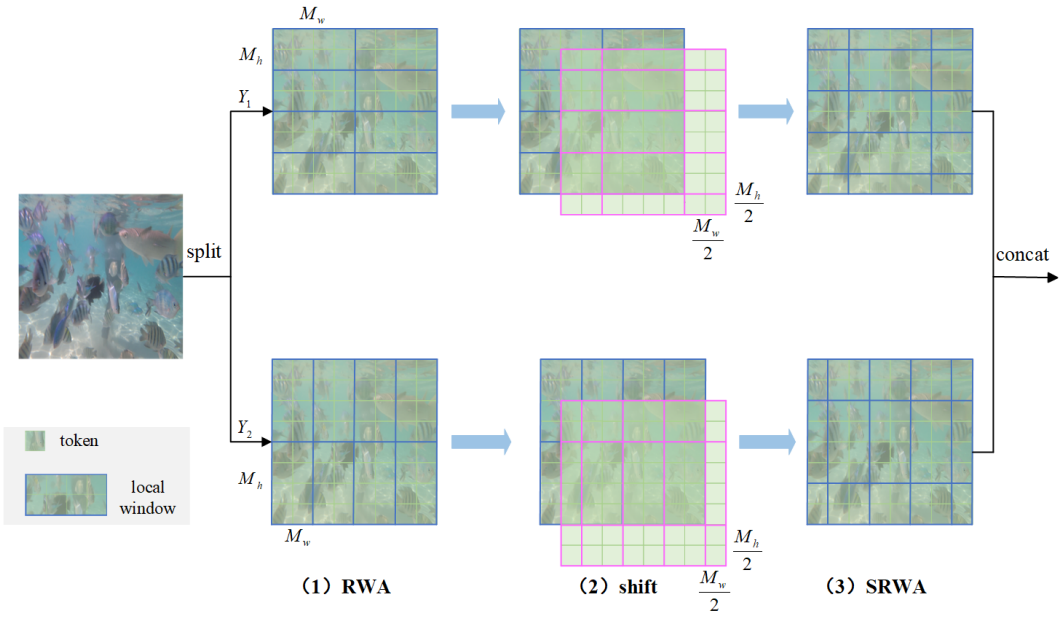

Figure 2. The window partition method of RWA and SRWA

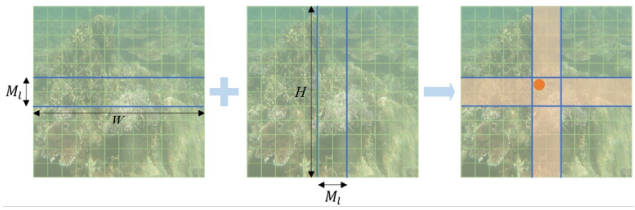

Figure 3. The attention area of the pixel in ARWA

Furthermore, one side of the rectangular window is extended to match the full length of the input feature map (either H or W), while the other side takes a smaller value, denoted as M_l. As illustrated in Fig. 3, the orange pixel can now interact with all the pixels along its horizontal and vertical axes, within the orange-shaded region. This type of attention calculation is referred to as Axial Rectangle Window Attention (ARWA). Similarly, a Shift Axial Rectangle Window Attention (SARWA) is added after ARWA, and they alternate.

In terms of computational cost, the standard Transformer computes the similarity between pixels over the entire feature map of size $H \times W \times C$, which results in a computational complexity proportional to the square of the feature map size, as shown in Equation (5). The Rectangle Window Attention (RWA) computes self-attention within a series of rectangular windows of size $M_h \times M_w \times C$, and the number of such windows is $\frac{H}{M_h} \times \frac{W}{M_w}$, leading to the computational complexity as given by Equation (6).

$$\begin{aligned} \Omega(MSA) &= 2(HW)^2C + 4HWC^2 \\ &= HWC \times (2HW + 4C) \end{aligned} \tag{5}$$

$$\begin{aligned} \Omega(RWA) &= (2(M_hM_w)^2 + 4M_hM_wC^2) \times \frac{H}{M_h} \times \frac{W}{M_w} \\ &= HWC \times (2M_hM_w + 4C) \end{aligned} \tag{6}$$

Similarly, the computational complexity for Axial Rectangle Window Attention (ARWA) is provided in Equation (7):

$$\Omega(ARWA) = HWC \times (HM_l + WM_L + 4C) \tag{7}$$

Although the computational cost of ARWA is higher than that of RWA, it is significantly reduced compared to the standard Transformer. Moreover, ARWA's window size changes flexibly with the feature map, making it more adaptable. Additionally, ARWA has a larger attention region, capable of capturing all information in both horizontal and vertical directions. Therefore, in the Multiple Attention Block (MAB), ARWA and SARWA are employed alternately.

3.4. Convolution-Enhanced Window Attention Calculation

Transformers are known for effectively capturing global dependencies, while Convolutional Neural Networks (CNNs) excel at local feature extraction and translation invariance, making them effective in capturing

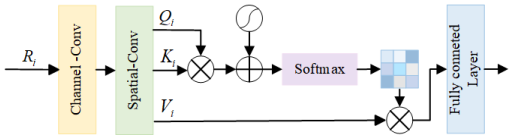

Figure 4. The algorithm of the window attention combined with the convolution

the two-dimensional local structure of images. Therefore, convolution operations are integrated into the calculation of rectangular window attention, resulting in Convolution-Axial Rectangle Window Attention (Conv-ARWA) and Convolution-Shift Axial Rectangle Window Attention (Conv-SARWA) to enhance local information within the windows and improve the model's ability to reconstruct image details.

In conventional window attention calculations [27], a fully connected layer is used to generate the Query, Key, and Value matrices from the input window feature map. In the proposed method, two convolutional layers are applied within the divided windows—one along the channel dimension and the other along the spatial dimension—to extract local features and generate the corresponding Query, Key, and Value matrices for each window. Fig. 4 illustrates the workflow of convolution-enhanced window attention calculation.

By incorporating convolution operations, the Conv-ARWA and Conv-SARWA modules can extract richer local details, thereby enhancing the ability of the model to effectively reconstruct detailed features in underwater images. This fusion helps in retaining essential information that might be overlooked by purely attention-based mechanisms, ultimately leading to improved image enhancement results.

After dividing the feature map into a series of rectangular windows according to the approach in Section 2.3, denote the horizontal rectangular windows obtained from the first $\frac{C}{2}$ channels of the feature map as $R_i \in R^{M_h \times M_w \times \frac{C}{2}} [i = 1, 2, ..., \frac{H \times W}{M_h \times M_w}, M_h < M_w]$.

For window R_i, a 1×1 convolution is first applied along the channel dimension to facilitate information interaction, and the number of channels is expanded by three times. Then, a 3×3 convolution is used to enhance the extraction of local spatial features, resulting in features $G_i \in R^{M_h \times M_w \times \frac{3C}{2}}$, as shown in Equation (8):

$$G_i = Conv_{3\times3}(Conv_{1\times1}(R_i)) \tag{8}$$

The dimensions of G_i are then reshaped to $M_h \times M_w \times \frac{3C}{2}$, and it is split evenly along the channel dimension to obtain the Query, Key, and Value matrices for the window, denoted as $Q_i \in R^{M_h \times M_w \times \frac{C}{2}}$, $K_i \in R^{M_h \times M_w \times \frac{C}{2}}$ and $V_i \in R^{M_h \times M_w \times \frac{C}{2}}$, respectively. The matrices Q_i, K_i and V_i are further divided into d heads along the channel dimension, with each head having a channel dimension of $D = C/2d$. The attention is computed for each head of the window in parallel, as shown in Equation (9):

$$\begin{aligned} Q_i &= [Q^1, Q^2, ..., Q^d], \\ K_i &= [K^1, K^2, ..., K^d], \\ V_i &= [V^1, V^2, ..., V^d] \end{aligned} \tag{9}$$

The attention calculation process for the m-th head of horizontal window R_i is represented by Equation (10):

$$\begin{aligned} X_i^m &= Attention(Q_i^m, K_i^m, V_i^m) \\ &= Softmax[\frac{Q_i^m (K_i^m)^T}{\sqrt{D}} + B]V_i^m, \\ m &= 1, 2, ..., d \end{aligned} \tag{10}$$

Where B represents the learnable positional encoding, and T represents matrix transpose. The attention results from all d heads for the horizontal window are concatenated along the channel dimension to obtain the attention output for the i-th horizontal window, denoted as $F_i \in R^{M_h \times M_w \times \frac{C}{2}} (M_h < M_w)$:

$$F_i = Concat(X_i^m), m = 1, 2, ..., d \tag{11}$$

Finally, the attention outputs from all $\frac{H \times W}{M_h \times M_w}$ horizontal windows are recombined in the original partitioned order to restore the size of the original feature map, obtaining the horizontal window attention output $F \in R^{H \times W \times \frac{C}{2}}$:

$$\begin{aligned} F &= WindowReverse(F_i), \\ i &= 1, 2, ..., \frac{H \times W}{M_h \times M_w} \end{aligned} \tag{12}$$

This process effectively extracts both local and global dependencies within each window and integrates them back to enhance the feature representation of the entire feature map.

For the vertical windows obtained from the remaining $\frac{C}{2}$ channels, attention is computed following the same steps as described above. Let the attention output for the i-th vertical window be denoted as $V - i \in R^{M_h \times M_w \times \frac{C}{2}} (M_h > M_w)$. The final attention output for all vertical windows, denoted as $V \in R^{H \times W \times \frac{C}{2}}$, is computed as follows:

$$V = WindowsReverse(Vi), i = 1, 2, ..., \frac{H \times W}{M_h \times M_w} \tag{13}$$

Next, the outputs F and V are concatenated along the channel dimension, and a fully connected layer is applied to obtain the final output of the rectangular window attention, denoted as $X_R \in R^{H \times W \times C}$:

$$X_R = Concat(F, V)W^o \tag{14}$$

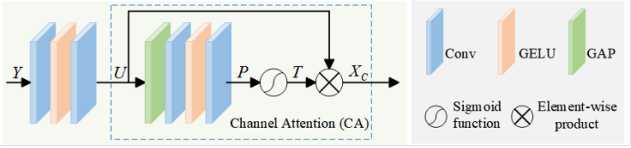

Figure 5. The structure of CEM

Where $W^o \in R^{C\times C}$ represents the linear mapping matrix. This step ensures that the information from both horizontal and vertical windows is fully integrated, capturing comprehensive spatial features and enriching the final feature representation.

3.5. Channel Enhanced Module

The channel dimension of an image contains rich color information, and enhancing the inter-channel associations can help in learning the color and intensity variations specific to underwater environments. In this work, a Channel Enhanced Module (CEM) is added in parallel with the rectangular window attention mechanism. The structure of CEM is shown in Fig. 5. The input feature Y is processed through two 3×3 convolutional layers with GELU activation functions to extract features $U \in R^{H\times W\times C}$, as represented by Equation (15):

$$U = Conv_{3\times3}(GELU(Conv_{3\times3}(Y))) \tag{15}$$

This operation aims to strengthen the correlation between different channels, capturing subtle variations in color that are critical for enhancing underwater images. By using pointwise convolutions, the module focuses solely on channel-wise interactions, without altering the spatial relationships, which effectively boosts the model's ability to reconstruct color and texture details.

Next, channel attention[32] is used to establish associations among different channels of the feature map, adaptively assigning channel weights. Specifically, the feature U is treated as consisting of C feature maps, each of size $H \times W$, denoted as $U = [v_1, v_2, ..., v_C]$. Global Average Pooling (GAP) is applied to capture the global information of each channel, as shown in Equation (16):

$$U_k^{'} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}U(i,j,k), k = 1, 2, ..., C \tag{16}$$

A 1×1 convolutional layer is used to reduce the number of channels to $\frac{C}{\gamma}$, where γ is the channel compression ratio, set to 4. Then, the GELU activation function is applied to enhance the nonlinearity of the features, followed by another $1times1$ convolutional layer to restore the number of channels back to C, as shown in Equation (17):

$$P = Conv_{1\times1}(GELU(Conv_{1\times1}(U^{'}))) \tag{17}$$

Subsequently, a gated activation function $Sigmoid = \frac{1}{1+e^{-x}}$ is applied to normalize the values of the feature map P between 0 and 1. Finally, element-wise multiplication is performed with the input feature U, resulting in the output feature of the Channel Enhanced Module X_C, as described in Equation (18):

$$X_C = U \cdot Sigmoid(P) = U \cdot T \tag{18}$$

3.6. Loss Function

A multi-task loss function is employed in this work, consisting of Charbonnier loss, gradient loss, and multi-scale structural similarity (MS-SSIM) loss. The weighted combination of these three loss components is given by Equation (19):

$$L = L_C + \lambda_1 \cdot L_{GP} + \lambda_2 \cdot L_{MS} \tag{19}$$

Where λ_1 and λ_2 are the balancing coefficients for the loss terms, with values set to 2 and 1, respectively. The Charbonnier loss aims to reduce the pixel-wise difference between the enhanced output image I_{out} and the reference image I_{gt}, as shown in Equation (20):

$$L_C = E_{I_{out}\sim P(g), I_{gt}\sim P(o)}\sqrt{(I_{gt} - I_{out})^2 + \epsilon^2} \tag{20}$$

Where $P(o)$ and $P(g)$ represent the distributions of the reconstructed enhanced image I_{out} and the reference image I_{gt}, respectively. ϵ is set to $1e^{-3}$ to prevent the gradient from becoming zero, thereby improving robustness to a certain extent. This loss ensures that the enhanced output is as close as possible to the reference, providing a strong foundation for pixel-level similarity and helping the model to effectively reduce artifacts and retain important details. Considering that Charbonnier loss lacks attention to the high-frequency information of the image, gradient loss [23] is used to emphasize the reconstruction of high-frequency details by constraining the difference between the enhanced image I_{out} and the reference image I_{gt} in terms of spatial gradients. The calculation is as follows:

$$I_{GP} = E_{\nabla I_{gt}\sim Q(g), \nabla I_{out}\sim Q(o)}(\nabla I_{gt} - \nabla I_{out}) \tag{21}$$

Where ∇I_{gt} and ∇I_{out} represent the gradient fields of the reference image I_{gt} and the enhanced output image I_{out}, respectively. $Q(g)$ and $Q(o)$ represent the distributions of the gradients in the x and y directions for both ∇I_{gt} and ∇I_{out} .

To further improve the perceptual quality of I_{out} for human viewers, multi-scale structural similarity (MS-SSIM) loss [27] is used to reduce the differences between I_{gt} and I_{out} in terms of luminance, contrast, and structure at multiple scales.

As shown in Fig. 6, multiple scales of the image are obtained by repeatedly applying Gaussian filtering and 2x

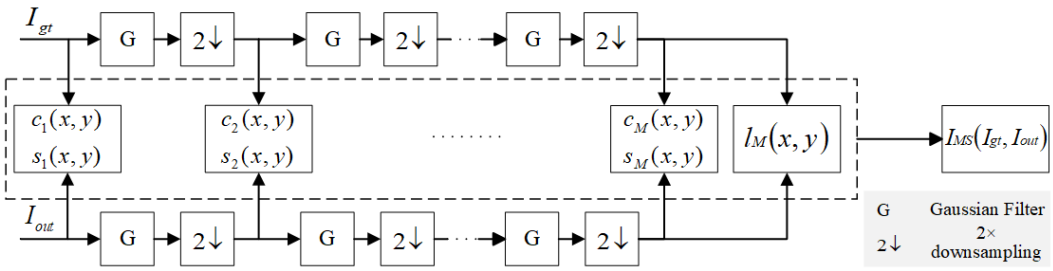

Figure 6. The algorithm of MS-SSIM

down-sampling to the original image. The original image is labeled as scale 1, and the image generated after $M-1$ iterations is labeled as scale M . The multi-scale structural similarity (MS-SSIM) is calculated as follows:

$$I_{MS}(I_{gt}, I_{out}) = [l_M(I_{gt}, I_{out})]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(I_{gt}, I_{out})]^{\beta_j} [s_j(I_{gt}, I_{out})]^{\gamma_j} \quad (22)$$

Where α^M, β_j and γ_j are the coefficients that adjust the importance of luminance, contrast, and structural components, respectively. $c_j(I_{gt}, I_{out})$ and $s_j(I_{gt}, I_{out})$ represent the contrast and structure differences at the j-th scale, and $l_M(I_{gt}, I_{out})$ represents the luminance difference at the M-th scale. The multi-scale structural similarity loss is computed as follows:

$$L_{MS} = 1 - I_{MS}(I_{gt}, I_{out}) \quad (23)$$

4. Experimental Results and Analysis

4.1. Experimental Setup

The experiments in this study were conducted using the Pytorch 1.11 framework on an NVIDIA V100 GPU with 24GB memory. The batch size for training was set to 8, and the number of training epochs was 800. The learning rate was set to 5×10^{-4} , and the Adam optimizer was used to train the model, with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The size of the underwater images used for both training and testing was uniformly adjusted to 256×256. The channel dimension C in the network was set to 64. In the PatchEmbed layer, the value of P was set to 2. In the Multiple Attention Block (MAB), six Rectangle Window Attention Layers (RWAL) and Shift Axial Rectangle Window Attention Layers (SRWAL) were used alternately, i.e., $N = 6$. The number of heads for multi-head self-attention in both horizontal and vertical rectangular windows was set to 4, i.e., $d = 4$.

4.2. Dataset

The experiments were conducted on the benchmark underwater image enhancement datasets UIEB [14] and UCCS [16]. The UIEB dataset contains 890 paired reference underwater images, with 800 pairs used as the training set, referred to as Train-800, and 90 pairs used as the testing set, referred to as Test-90. Additionally, UIEB includes 60 unpaired original underwater images, referred to as the Challenge Set (C60). The UCCS dataset contains 300 original underwater images without ground truth labels, consisting of 100 images each from three different color cast environments: blue, green, and blue-green. These datasets provide a diverse range of underwater conditions, allowing comprehensive evaluation of the model's ability to handle various underwater image enhancement challenges, including different lighting conditions and color distortions.

4.3. Evaluation Metrics

The evaluation metrics for underwater image enhancement in this study are divided into two categories: reference-based and non-reference. For underwater images with ground truth labels, reference-based metrics are used: Peak Signal-to-Noise Ratio (PSNR) [11] and Structural Similarity (SSIM) [26]. PSNR measures the pixel-wise difference between the enhanced image and the reference image, indicating the overall reconstruction accuracy. SSIM measures the contrast, luminance, and structural similarity between images, which aligns better with human visual perception. Higher values of PSNR and SSIM indicate a higher similarity between the enhanced image and the reference image in terms of content. For underwater images without ground truth labels, non-reference metrics are used: Underwater Image Quality Measure (UIQM) [18] and Underwater Color Image Quality Evaluation (UCIQE) [28]. UIQM is defined as a linear combination of Underwater Image Colorfulness Measure (UICM), Underwater Image Sharpness Measure (UISM), and Underwater Image Contrast Measure (UIConM). UCIQE considers chroma, saturation, and contrast in underwater images. Higher values of UIQM and UCIQE indicate better balance in terms of colorfulness, contrast, and overall visual quality of the enhanced image. These metrics provide a comprehensive assessment of the model's performance in both objective accuracy and perceptual quality, allowing for a detailed evalu-

ation of the enhancement results under different conditions and datasets.

4.4. Ablation Study

An ablation study was conducted on the UIEB dataset, with the model trained on Train-800 and tested on Test-90. To validate the contribution of the proposed modules to underwater image enhancement, the performance of three different network structures was evaluated.

• Model 1 integrates Axial Rectangle Window Attention (ARWA) and Shift Axial Rectangle Window Attention (SARWA) into the U-Net architecture.

• Model 2 adds convolution operations to the rectangular window attention mechanism, replacing ARWA and SARWA in Model 1 with Conv-ARWA and Conv-SARWA.

• Model 3 builds upon Model 2 by incorporating the Channel Enhanced Module (CEM), forming the final network, MAAU-UIE.

The experimental results are shown in Table 1. Compared to the baseline U-Net model, Model 1 achieved significant improvements in PSNR and SSIM, with increases of 5.629 and 0.084, respectively. This improvement is due to the limited global information modeling capability of the ResBlock, whereas Model 1 alternates between ARWA and SARWA, effectively combining local context extraction with global feature representation. Model 2, compared to Model 1, further increased PSNR by 0.250 and SSIM by 0.001, demonstrating that adding convolutional layers enhances the window attention mechanism's ability to learn spatial local information, thereby improving the model's detail enhancement capability. Model 3 achieved the best performance, indicating that integrating the Channel Enhanced Module allows for better utilization of global information, leading to improved reconstruction of color, brightness, and other variations. The combination of attention and channel enhancement contributed to a comprehensive enhancement of the underwater image quality.

Table 2 discusses the impact of different height (M_h) and width (M_w) combinations of the rectangular windows on the network performance. Based on Model 1, the height and width of the rectangular windows were varied. Compared to the first two rows with square window attention, Rectangle Window Attention (RWA) demonstrated significant performance advantages. Using Axial Rectangle Window Attention (ARWA) further improved the metrics, indicating that extending the longer side of the rectangular window to match the feature map's length better aggregates information along that direction. Additionally, increasing the shorter side of the axial rectangular window (M_l) allows each token to interact with more tokens, enhancing the extraction of global information. By comparing the last two rows of Table 2, it is observed that when M_l increases from 4 to 8, there is no significant performance improvement. Moreover, as indicated by Equation (7), increasing $M - l$ leads to higher computational cost. Therefore, considering the balance between image reconstruction performance and computational efficiency,M_l was set to 4.

Table 3 compares the effect of different values of α in the Channel Enhanced Module (CEM) on the performance of the MAAU-UIE network. Setting $\alpha = 0$ indicates that the channel enhancement module is not used. Adding the channel enhancement module ($\alpha > 0$) slightly improved PSNR and SSIM. When $\alpha = 0.1$, the model achieved optimal performance, indicating that adjusting the balance between channel information and spatial information fusion can further optimize the model's performance. The results highlight the importance of carefully choosing hyperparameters for the attention mechanisms and channel enhancement module to achieve a good balance between enhancement effectiveness and computational efficiency.

Table 4 compares the results of training the MAAU-UIE network using different combinations of the three loss components. When gradient loss (L_{GP}) or multi-scale structural similarity loss (L_{MS}) was added individually on top of the Charbonnier loss (L_C), both PSNR and SSIM showed significant improvements. When all three loss components were used together, both metrics reached their optimal values. This indicates that reducing the difference between the enhanced image and the reference image in the high-frequency gradient domain, as well as enhancing their structural consistency, contributes significantly to improving the visual quality of the output image. The combination of these losses effectively ensures that the enhanced images not only closely match the reference in pixel values but also retain important textural and structural information, ultimately leading to a more visually appealing enhancement result.

4.5. Quantitative Comparison

To validate the superiority of proposed method, various objective metrics were compared between the proposed approach and state-of-the-art underwater image enhancement methods. The models were trained on the UIEB training set (Train-800), and performance was compared on its testing set (Test-90) using PSNR and SSIM metrics. Additionally, for non-reference test sets (C60 from UIEB and UCCS), UIQM and UCIQE metrics were used for comparison.

Table 6 compares the UIQM and UCIQE metrics of various methods on C60 and UCCS datasets, respectively. According to Tables 6, the proposed method outperformed other methods on the UIQM metric for C60 and the UCIQE metric for UCCS, surpassing the second-best method by 0.004 and 0.003, respectively. These results indicate that the proposed method achieves superior sharpness and contrast, and exhibits strong robustness when processing underwater images with three different color casts in the UCCS dataset.

Model	U-Net(ResBlock)	ARWA & SARWA	Conv-ARWA & Conv-SARWA	CEM	Test-90	
					PSNR	SSIM
Base Model	✓				18.472	0.832
Model1		✓			24.101	0.916
Model2			✓		24.351	0.917
Model3			✓	✓	24.467	0.920

Table 1. Comparison of different obfuscations in terms of their transformation capabilities

	(M_h, M_w)		Test-90	
	First $C/2$ channel	Last $C/2$ channel	PSNR	SSIM
Square Window	(4,4)	(4,4)	22.120	0.899
	(8,8)	(8,8)	22.866	0.903
Rectangle Window	(2,4)	(4,2)	23.029	0.907
	(4,16)	(16,4)	23.154	0.909
	(8,16)	(16,8)	23.226	0.915
Axial Window	(2,W)	(H,2)	24.037	0.914
	(4,W)	(H,4)	24.101	0.916
	(8,W)	(H,8)	24.102	0.916

Table 2. The effects of different widths and heights of the rectangle window

α	Test-90	
	PSNR	SSIM
0	24.351	0.917
0.001	24.354	0.917
0.01	24.412	0.918
0.1	**24.467**	**0.920**
1	24.361	0.919

Table 3. Effects of the fusion factor XXX in CEM

Loss	Test-90	
	PSNR	SSIM
L_C	23.894	0.898
L_C, L_{MS}	24.210	0.911
L_C, L_{GP}	24.415	0.916
L_C, L_{GP}, L_{MS}	**24.467**	**0.920**

Table 4. Effects of different loss functions

Method	Venue	Test-90	
		PSNR	SSIM
Ucolor[28]	TIP 21	21.093	0.872
URSCT[14]	TGRS 22	22.720	0.910
FiveA+[10]	BMVC 23	23.061	0.911
U-shape[16]	TIP 23	22.910	0.910
PuGAN[17]	TIP 23	21.670	-
Semi-UIR[33]	CVPR 23	23.590	0.901
X-CAUNET[20]	ICASSP 24	24.121	0.871
PixMamba[29]	arXiv 24	23.587	**0.921**
Ours	-	**24.467**	0.920

Table 5. The performance indicators of different methods on the test set Test-90 of UIEB

This highlights the method's effectiveness in enhancing diverse underwater images and providing high visual quality, especially in challenging color-biased conditions.

Table 6 also compares the performance differences between our model and the current state-of-the-art models. From the metrics of parameters and FLOPs, it can be seen that our model outperforms most of the current state-of-the-art models.

5. Qualitative Comparison

To further verify the superiority of the proposed method, the visual enhancement effects on Test-90, C60, and UCCS were compared against mainstream underwater image enhancement algorithms, including URSCT [20], U-shape [12], PuGAN [18], X-CAUNET [3], and Pixmaba [22].

Fig. 7 illustrates the enhancement results of different methods on Test-90. For example, URSCT shows issues of insufficient contrast and blurry details in the enhancement results of rows 1, 2, and 3. X-CAUNET produces an image with an overall reddish color cast in row 1, while the result in row 4 has a greenish tint and lacks clarity. Pixmaba produces dark images in row 4 and introduces yellow artifacts in the enhanced image in row 6.

The proposed method effectively avoids these issues, achieving higher contrast while enhancing detail information and improving color realism. This demonstrates the robustness and effectiveness of the method in providing visually superior enhancement results compared to existing approaches, especially in challenging underwater environments.

Fig. 8 visualizes the output results of different methods on the C60 dataset. PuGAN shows issues of local over-enhancement and poor color balance in rows 1 and 2. X-CAUNET results in darkened areas in the enhanced images of rows 1 and 3. Pixmamba produces images with low con-

Table 6. The performance indicators of different methods on the challenge set C60 of UIEB and UCCS

Method	Venue	C60		UCCS		Params	FLOPs
		UIQM	UCIQE	UIQM	UCIQE		
Ucolor[12]	TIP 21	2.482	0.553	3.019	0.55	157.4M	34.68G
URSCT[21]	TGRS 22	2.642	0.543	2.947	0.544	11.41M	18.11G
MFEF[33]	EAAI 23	2.652	0.566	2.977	0.55	61.86M	26.52G
PuGAN[4]	TIP 23	2.652	0.566	2.977	0.53	95.66M	72.05G
Semi-UIR[7]	CVPR 23	2.667	0.574	**3.079**	0.55	**1.65M**	36.44G
X-CAUNET[20]	ICASSP 24	2.683	0.564	2.922	0.541	31.78M	261.48G
PixMamba[15]	arXiv 24	2.868	**0.586**	3.053	0.561	8.68M	**7.60G**
Ours	-	**2.872**	0.572	3.066	**0.564**	8.52M	9.84G

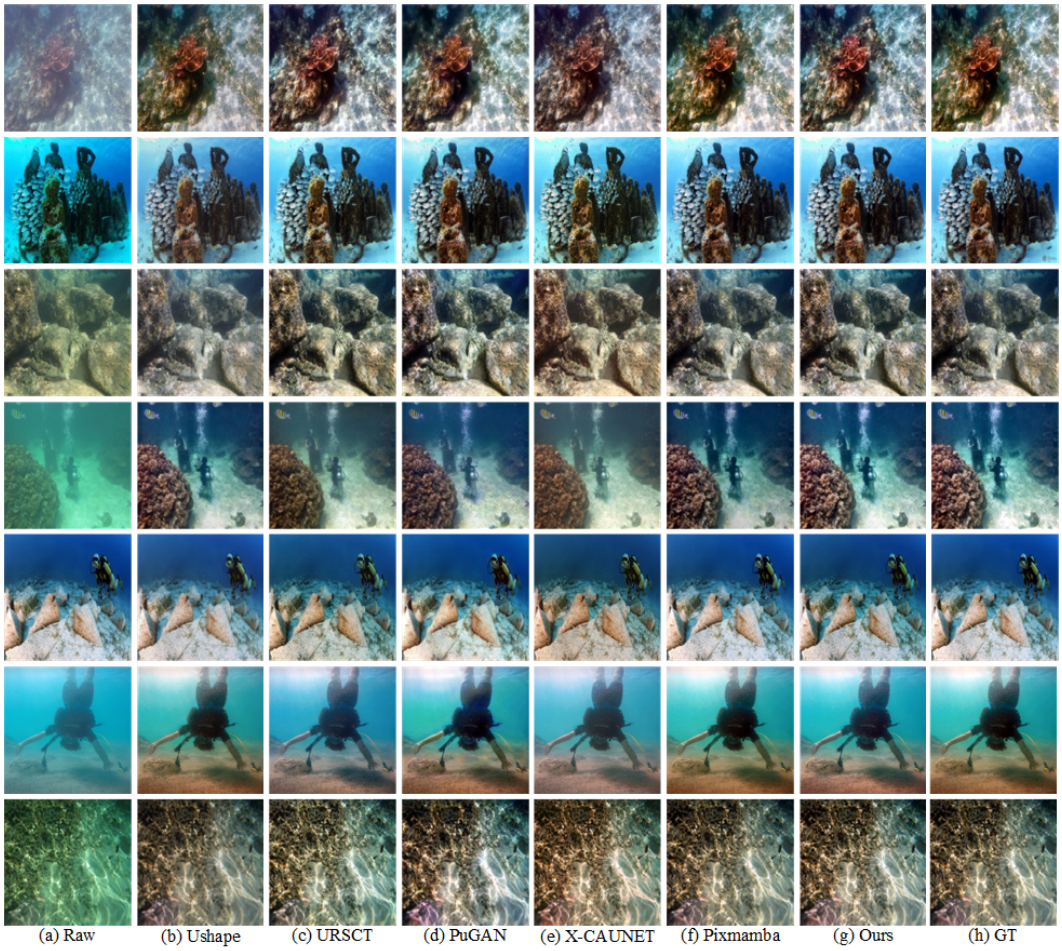

Figure 7. Visual comparison of differernt methods on the test set of UIEB named Test-90

trast in rows 3 and 5. In contrast, the proposed method improves image contrast while avoiding issues of over-bright or overly dark regions.

Fig. 9 shows the enhancement results of different methods on the UCCS dataset. The first two rows, the middle two rows, and the last three rows respectively show visual results from underwater environments with blue, green, and blue-green color casts in the UCCS dataset. It can be observed that the proposed method effectively removes the blue-green background and reduces color bias, resulting in images that are more consistent with human visual perception. These visual comparisons demonstrate the capability of the proposed method to handle different color cast environments effectively, ensuring balanced color distribution, improved contrast, and enhanced overall visual quality, thereby making the enhanced images more natural and

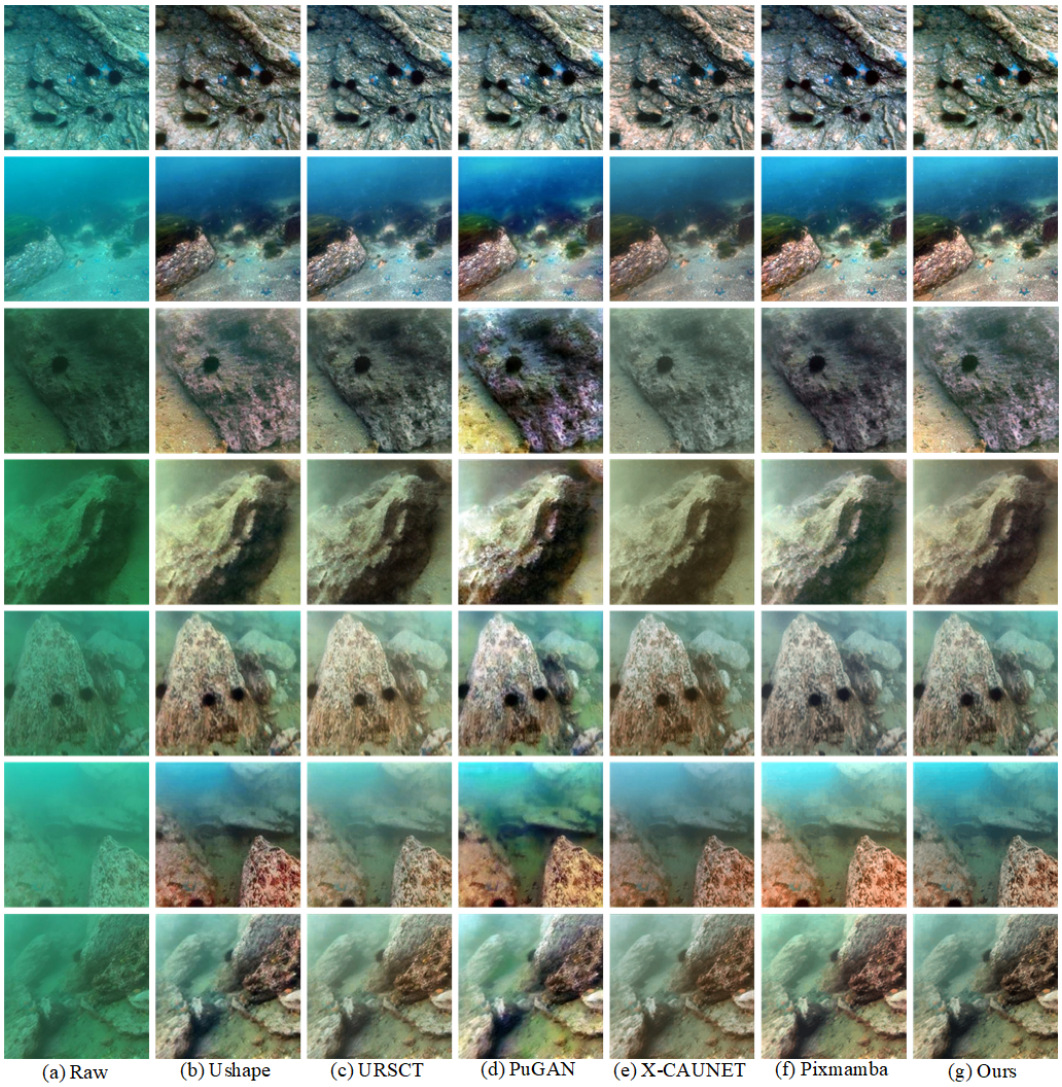

Figure 8. Visual comparison of differernt methods on the challenge set of UIEB named C60

appealing to the human eye.

6. Conclusion

To address issues of color distortion, low contrast, and blurred details in underwater images, this paper proposed a Multiple Attention Aggregation U-Net for Underwater Image Enhancement (MAAU-UIE). The model combines Axial Rectangle Window Attention (ARWA) and Shift Axial Rectangle Window Attention (SARWA), leveraging their respective advantages in extracting local features and enhancing global information. The convolutional operations further enhance the ability of window attention to learn local details, while the Channel Enhanced Module (CEM) allows the network to focus more on important channel information. Additionally, gradient loss and multi-scale structural similarity loss are used to optimize the network's ability to recover edges and structures. Experimental results demonstrate that the proposed improvements effectively enhance the performance of underwater image enhancement. Compared to state-of-the-art methods, the proposed approach achieves superior results in PSNR, SSIM, UIQM, and UCIQE metrics, while also providing reconstructed images with realistic colors, high sharpness, and good contrast. Currently, the network's foundational modules are designed based on the Transformer architecture. Future work will focus on exploring more advanced Mamba architectures to further improve performance metrics and inference efficiency.

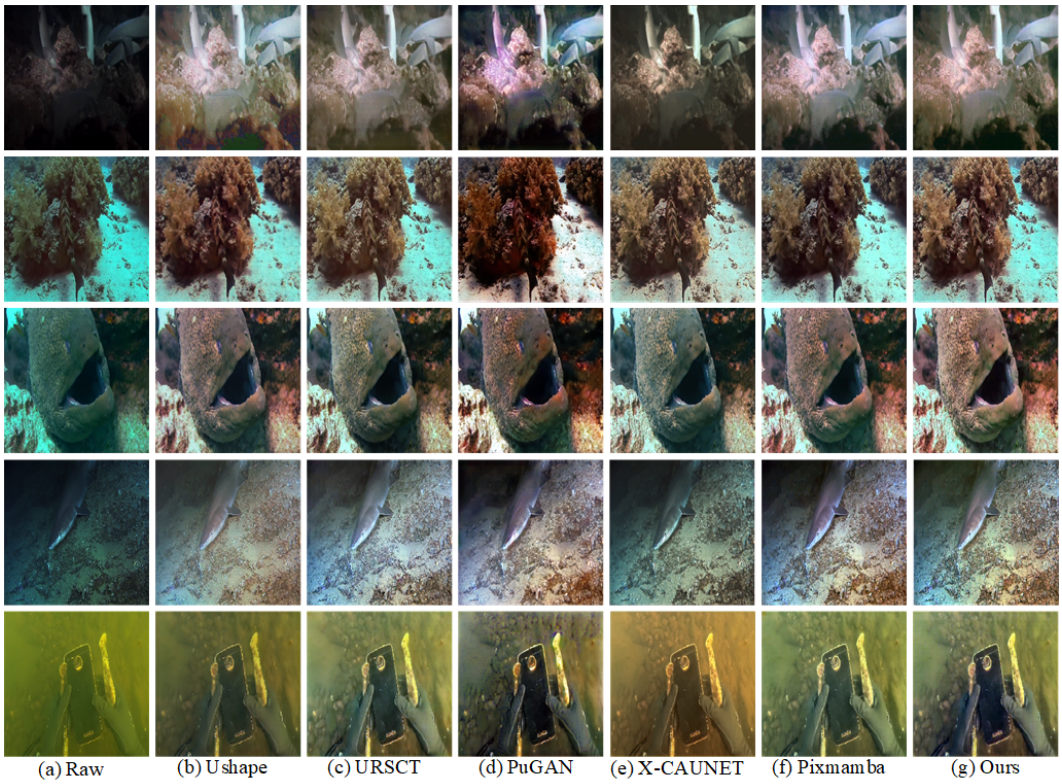

Figure 9. Visual comparison of different methods on the UCCS

References

[1] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert. Color balance and fusion for underwater image enhancement. *IEEE Transactions on image processing*, 27(1):379–393, 2017. 1

[2] D. Berman, D. Levy, S. Avidan, and T. Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2822–2837, 2020. 1, 2

[3] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490, 2022. 4

[4] R. Cong, W. Yang, W. Zhang, C. Li, C.-L. Guo, Q. Huang, and S. Kwong. Pugan: Physical model-guided underwater image enhancement using gan with dual-discriminators. *IEEE Transactions on Image Processing*, 32:4472–4485, 2023. 2, 11

[5] Z. Fu, W. Wang, Y. Huang, X. Ding, and K.-K. Ma. Uncertainty inspired underwater image enhancement. In *European conference on computer vision*, pages 465–482. Springer, 2022. 1

[6] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[7] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18145–18155, 2023. 11

[8] Z. Huang, J. Li, Z. Hua, and L. Fan. Underwater image enhancement via adaptive group attention-based multiscale cascade transformer. *IEEE Transactions on Instrumentation and Measurement*, 71:1–18, 2022. 2

[9] M. J. Islam, Y. Xia, and J. Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020. 2

[10] J. Jiang, T. Ye, J. Bai, S. Chen, W. Chai, S. Jun, Y. Liu, and E. Chen. Five a+ network: You only need 9k parameters for underwater image enhancement. arxiv 2023. *arXiv preprint arXiv:2305.08824*. 10

[11] J. Korhonen and J. You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, pages 37–38. IEEE, 2012. 8

[12] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Transactions on Image Processing*, 30:4985–5000, 2021. 1, 11

[13] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao. An underwater image enhancement benchmark

dataset and beyond. *IEEE transactions on image processing*, 29:4376–4389, 2019. 1

[14] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE transactions on image processing*, 29:4376–4389, 2019. 2, 8, 10

[15] W.-T. Lin, Y.-X. Lin, J.-W. Chen, and K.-L. Hua. Pixmamba: Leveraging state space models in a dual-level architecture for underwater image enhancement. In *Proceedings of the Asian Conference on Computer Vision*, pages 3622–3637, 2024. 2, 11

[16] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE transactions on circuits and systems for video technology*, 30(12):4861–4875, 2020. 2, 8, 10

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 4, 10

[18] K. Panetta, C. Gao, and S. Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2015. 8

[19] L. Peng, C. Zhu, and L. Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 32:3066–3079, 2023. 2

[20] A. Pramanick, S. Sarma, and A. Sur. X-caunet cross-color channel attention with underwater image-enhancing transformer. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3550–3554. IEEE, 2024. 1, 10, 11

[21] T. Ren, H. Xu, G. Jiang, M. Yu, X. Zhang, B. Wang, and T. Luo. Reinforced swin-convs transformer for simultaneous underwater sensing scene image enhancement and super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 2, 11

[22] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[23] J. Sun, Z. Xu, and H.-Y. Shum. Gradient profile prior and its applications in image super-resolution and enhancement. *IEEE Transactions on Image Processing*, 20(6):1529–1542, 2010. 7

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[25] R. Wang, S. Wang, Y. Wang, L. Cheng, and M. Tan. Development and motion control of biomimetic underwater robots: A survey. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(2):833–844, 2020. 1

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[27] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 6, 7

[28] M. Yang and A. Sowmya. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24(12):6062–6071, 2015. 8, 10

[29] X. Ye, H. Xu, X. Ji, and R. Xu. Underwater image enhancement using stacked generative adversarial networks. In *Pacific Rim Conference on Multimedia*, pages 514–524. Springer, 2018. 2, 10

[30] F. Yu, B. He, and J.-X. Liu. Underwater targets recognition based on multiple auvs cooperative via recurrent transfer-adaptive learning (rtal). *IEEE Transactions on Vehicular Technology*, 72(2):1574–1585, 2022. 1

[31] D. Zhang, C. Wu, J. Zhou, W. Zhang, C. Li, and Z. Lin. Hierarchical attention aggregation with multi-resolution feature learning for gan-based underwater image enhancement. *Engineering Applications of Artificial Intelligence*, 125:106743, 2023. 2

[32] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 7

[33] J. Zhou, J. Sun, W. Zhang, and Z. Lin. Multi-view underwater image enhancement method via embedded fusion mechanism. *Engineering applications of artificial intelligence*, 121:105946, 2023. 10, 11

[34] P. Zhuang, J. Wu, F. Porikli, and C. Li. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Transactions on Image Processing*, 31:5442–5455, 2022. 1, 2