Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Underwater image enhancement method based on a cross attention mechanism

Sunhan Xu
Beijing Union University, Smart City College,China
Jinhua Wang (xxtwangjinhua@buu.edu.cn)
Beijing Union University, Smart City College,China
Ning He
Beijing Union University, Smart City College,China
Xin Hu
Beijing Union University, Smart City College,China
Fengxi Sun
Beijing Union University, Smart City College,China

Research Article

Keywords: Underwater image enhancement methods, U-Net, Cross attention Transformer, Dynamic enhancement module, Hybrid loss function

Posted Date: September 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3285291/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on January 19th, 2024. See the published version at https://doi.org/10.1007/s00530-023-01224-5.

Underwater image enhancement method based on a cross attention mechanism

Sunhan Xu[1], Jinhua Wang[1*], Ning He[1], Xin Hu[1], Fengxi Sun[1]

[1*]Smart City College, Beijing Union University, Chaoyang, Beijing, 100020, Beijing, China.

*Corresponding author(s). E-mail(s): xxtwangjinhua@buu.edu.cn;
Contributing authors: 20221081210206@buu.edu.cn;
xxthening@buu.edu.cn; 20211081210210@buu.edu.cn;
20221081210201@buu.edu.cn;

Abstract

Underwater image enhancement is a technique that improves the quality of underwater images, which makes them clearer and more realistic. However, because of the complexity of underwater environments, underwater image enhancement faces many challenges, such as the variation in underwater optical properties as well as low contrast, low brightness, and color distortion in underwater images. To extract underwater image features more effectively, this paper proposes an underwater image enhancement algorithm called Cross Attention-based Underwater Image Enhancement (CAUIE). The algorithm combines cross attention and dynamic enhancement modules to build a U-Net model. Cross attention uses a self-attention mechanism to capture the local and global information of underwater images, thus enhancing the semantic representation of the images. The dynamic enhancement module, by contrast, dynamically adjusts the enhancement parameters according to different regions of the image to acquire detail information. In addition, this paper introduces a contrastive regularization loss to construct a hybrid loss function for guiding the training and optimization of the model. The experimental results show that the proposed algorithm outperforms the comparison algorithm in both subjective visual and objective evaluation criteria. Moreover, the proposed model obtains PSNR and SSIM results of 34.86 dB and 0.996, respectively, increasing the results of the previous model by 7.97 dB and 0.099, which illustrates that the proposed algorithm can solve the color distortion problem and recover the contrast and clarity of underwater images.

Keywords: Underwater image enhancement methods,U-Net,Cross attention Transformer,Dynamic enhancement module,Hybrid loss function

1 Introduction

Because the underwater environment is complex and dynamic, underwater image processing techniques face great challenges. Underwater images are affected by many factors, such as light propagation, absorption and scattering as well as the concentration of suspended particles in water, all of which have a huge impact on image quality, making underwater image processing extremely difficult. Therefore, efficient and reliable underwater image enhancement techniques are highly important for deep exploration of the ocean.

Underwater image enhancement techniques improve three main aspects of the visual quality of underwater images: visibility, color cast, and contrast. Previous underwater image enhancement methods have used convolution neural networks (CNN) and generative adversarial networks (GAN) [1] to enhance underwater images . However, a CNN is mainly used to extract local information, and it is difficult for a CNN to extract global information. Lan et al. [2] proposed QACG that has introduced attention mechanisms to extract global information, but the improvement in the final results still needs to be increased. By contrast, Vision Transformer (ViT) [3] are adept at extracting global information from images, although they tend to overlook local information. Additionally, the series of Transformers using attention mechanisms often suffer from excessive computation and a large number of parameters. To solve this problem, this paper proposes an underwater image enhancement model based on U-Net [4], which combines the cross attention Transformer (CAT) [5] and dynamic enhancement modules to effectively extract features from underwater images and perform adaptive enhancement. In the CAT framework, a large kernel attention (LKA) mechanism [6] is added that uses dilated convolution to simulate the attention mechanism, which greatly reduces the amount of computation and parameters of the model. To better optimize the model, a hybrid loss function is designed that includes pixel loss, contrast regularization (CR) loss, and structural similarity (SSIM) loss to guide the training and optimization of the model.

The experimental results show that the algorithm proposed in this paper is superior to current mainstream algorithms in terms of the PSNR and SSIM metrics, especially on the HICRD [7], where significant improvements were achieved. With respect to the results of the best comparison model, the PSNR was improved by 7.97 dB and the SSIM was improved by 0.099. On the UIEBD [8], a PSNR of 22.12 dB and SSIM of 0.889 were obtained, which are better than the results obtained by the comparison algorithms.

2 Related Work

Recent years have witnessed two major approaches for underwater image enhancement: methods based on prior knowledge [9–14] and methods based on deep learning [7, 15–20].

2.1 Underwater Image Enhancement Based on Prior Knowledge

In the field of underwater image enhancement, methods based on prior knowledge have been widely applied. These methods estimate the parameters of underwater imaging models through prior assumptions.

He et al. [9] proposed an image dehazing method based on the dark channel prior (DCP), which can directly estimate the thickness of haze and recover high-quality, haze-free images. For underwater images, Drews et al. [10] improved DCP and proposed an algorithm called UDCP, which considers the blue and green color channels as the source of underwater visual information, obtaining significant improvements when compared with the original DCP [9]. Peng et al. [11] proposed an underwater image depth estimation method called IBLA, which estimates the degradation model based on image blurriness and light absorption and uses this model to restore the image. However, for images with distinct bright and dark regions, excessive contrast stretching can lead to overly extreme bright and dark areas, resulting in loss of details. Akkaynak et al. [12] proposed a new underwater image restoration model that better conforms to the physical characteristics of underwater light and can more accurately estimate the depth and background light of underwater scenes to improve the quality of restored underwater images. Based on this, in 2019, they further proposed an underwater image restoration method called Sea-thru [13], which uses RGBD images and rectified underwater images to form a model for restoring underwater image colors.

Cao et al. [14] proposed an underwater image restoration method that uses deep networks to estimate background light and scene depth. They designed two neural network structures, one to estimate background light and the other to estimate scene depth, and then applied them to the underwater image formation model to restore underwater images. This method can handle different water types and lighting conditions, and has clear robustness and adaptability.

In summary, in existing underwater image enhancement methods based on prior knowledge, accurately estimating the parameters of underwater image formation models still remains challenging because of the complexity and diversity of underwater scenes.

2.2 Underwater Image Enhancement Based on Deep Learning

In recent years, deep neural networks have achieved tremendous success in computer vision, which has inspired researchers to attempt to improve the performance of underwater image enhancement using deep learning methods.

Li et al. [16] proposed a method called WaterGAN that uses GAN to generate realistic underwater images from aerial images and depth maps for color correction of monocular underwater images. Li et al. [17] proposed a deep underwater image and video enhancement method based on underwater scene priors. They synthesized underwater image training data using these underwater scene priors and designed a lightweight CNN model to enhance images for each underwater scene type to directly reconstruct clear underwater images and improve contrast, saturation, and brightness. However, this method suffers from over-compensation and insufficient generalization

capabilities. Han et al. [15] proposed an underwater image restoration model called CWR, which utilizes contrastive learning and GAN to maximize the mutual information in the original and restored images. This helps enhance the clarity and color restoration of underwater images while preserving textures and structures. Li et al. [8] constructed an underwater image enhancement benchmark dataset called the Underwater Image Enhancement Benchmark (UIEB), and proposed a deep underwater image enhancement network called WaterNet based on underwater scene priors. Li et al. [18] proposed an underwater image enhancement network model called Ucolor, which uses features from multiple color spaces and the guidance of underwater image formation models to improve the color and contrast of underwater images. Kar et al. [20] proposed a zero-shot image enhancement method based on Koschmieder's light scattering model, which restores images by controlling the perturbations in the model without learning scene-specific or distortion-specific knowledge. This method has achieved promising results in image dehazing, underwater image restoration, and similar tasks. Fu et al. [19] proposed an uncertainty-inspired underwater image enhancement model called PUIENet, which establishes an enhancement distribution of underwater image formation models using probabilistic networks, samples multiple enhancement predictions from it, and then predicts deterministic results through consensus process. This method can handle the uncertainty in labeled reference images and improves the clarity and color restoration of underwater images while preserving structure and texture details.

However, despite the effectiveness of existing learning-based underwater image enhancement methods in improving the visual results, the complexity and diversity of underwater images means that existing models still struggle to meet practical needs. Further research and exploration are still required in this field.

2.3 ViT

In recent years, the ViT [3] has gradually replaced CNNs in computer vision applications. A ViT divides images into patches and uses self-attention to capture global features. However, a ViT incurs excessive computational cost and cannot extract local information from images. To address this, the Swin Transformer [21] adopts shifted windows to capture more features with linear computational cost. Dehazeformer [22] combines a Swin Transformer [21] and U-Net[4], and incorporates convolutions into the Swin Transformer to better extract image features, achieving great results in image dehazing. Lin et al. [5] proposed a new attention mechanism called cross attention and constructed a hierarchical network model called CAT. The key idea is to apply attention alternately within and between image patches to reduce computational cost and capture both local and global information.

Because the ViT [3] is complex, it is highly demanding regarding training resources and time. In recent years, many researchers have proposed improved versions to address this issue. For example, the pyramid vision Transformer [23] uses strided convolutions to reduce the number of times the attention mechanism needs to be computed, thus lowering the use of computational resources. MobileViT [24] combines a ViT architecture with a CNN to create a lightweight design. Many researchers also use MLPs to approximate the attention mechanism, as in LKA [6]. Yu et al. [25] proposed

Poolformer and Metaformer, which are extremely lightweight ViTs that replace the self-attention mechanism with pooling layers, but they inevitably suffer some decrease in accuracy.

Inspired by the Dehazeformer[22] architecture and CAT's capability to extract global and local image information, they are combined in the proposed method. However, the self-attention[26] in CAT[5] has a large number of parameters and high computational cost, requiring considerable computational resources, which means that the size of the CAT model is large. To reduce the number of parameters and computational cost, the method proposed in this paper replaces the self-attention[26] in the original CAT[5] with LKA[6].

3 Method

The proposed underwater image enhancement algorithm, Cross Attention-based Underwater Image Enhancement (CAUIE), consists of the following three types of modules: (1) convolution modules , (2) cross large kernel attention Transformer (CLKAT), and (3) dynamic feature enhancement (DFE) module. The convolution module includes common operations such as convolution, up-sampling, and down-sampling. To improve the performance of underwater image enhancement, the model is trained with a combination of three losses, each playing a different role.

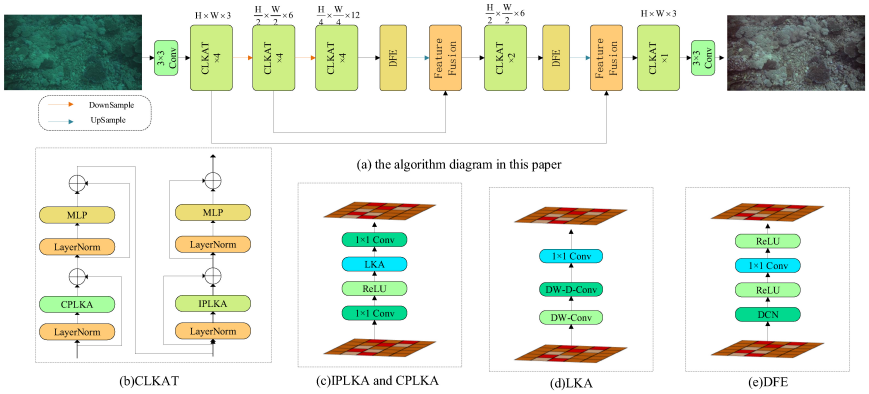

Fig. 1 Diagram of the proposed algorithm. CAUIE is a modified 5-layer U-Net in which the convolution modules are replaced by CLKAT and DFE is applied before up-sampling. (b), (c), and (d) Details of CLKAT, which primarily involves the use of the modified LKA instead of the original self-attention in CAT. (e) Details of DFE. The input size is $H \times W$, and the feature map size at each stage is indicated in the diagram.

3.1 Underwater Image Enhancement Network

CAUIE adopts a U-Net[4] structure consisting of feature extraction and image reconstruction. The feature extraction process extracts feature representations, whereas the

reconstruction process reconstructs the image. Fig.1 shows the overall framework of the proposed method. First, the input image goes through a 3×3 convolution to extract low-level features. The extracted features are then fed into the CLKAT to obtain additional feature information, which is processed by the DFE. After two up-sampling and down-sampling processes, the feature map size is adjusted accordingly. Finally, the image is reconstructed using a 3×3 convolution module after the CLKAT. Throughout the process, feature fusion is used to fuse the branches after the CLKAT, DFE module, and the original image branch to incorporate low-level features and hierarchical CLKAT features. The use of DFE improves the model's ability to extract image detail information, whereas CLKAT effectively captures and integrates the local and global information in the image.

3.1.1 Cross Large Kernel Attention

The proposed CLKAT was inspired by the CAT[5] and LKA[6]. The CAT consists of two parts: intra-position self-attention (IPSA) and cross-position self-attention (CPSA). Because of the large number of parameters and computational overhead of self-attention[26], we were inspired by the recent practice of using MLPs to replace self-attention to substitute the self-attention[26] in CAT[5] with LKA [6] to create the intra-position LKA (IPLKA) and cross-position LKA (CPLKA). The key idea of LKA [6] is to replace the high-parameter and high-computation self-attention operation[26] with convolution operations, such as dilated convolution and depth-wise convolution [27], that have smaller numbers of parameter and computational overhead. In addition, Song et al. [22] have experimentally and theoretically proven that ReLU is more suitable than GELU for low-level computer vision tasks. Therefore, the method proposed in this paper improves on the attention in VAN [6] by replacing the GELU in the original attention with ReLU. The modified architectures of CLKAT, the attention used in the proposed method, and LKA [6] are illustrated in Fig. 1(b), (c), and (d), respectively. IPLKA extracts local information from the feature maps, and CPLKA extracts global information from the feature maps after local feature extraction. The original CAT [5] has three sequential layers, IPSA, CPSA, and IPSA. After considering multiple model designs and ablation experiments, the final design of CLKAT has two layers, as shown in Fig.1(b). CLKAT first performs CPLKA to extract global features from the image, followed by IPLKA to extract local information from the feature maps. CLKAT can employ attention between image patches divided from single-channel feature maps to capture global information. The depth of CLKAT at each layer is 4, 4, 4, 2, and 1.

3.1.2 Feature Fusion Module

The feature fusion in the proposed method is a channel attention-based method that can automatically select the most important channels and fuse the original features and cross attention features based on the importance of each channel to the model output. It mainly consists of concatenation, pooling, and softmax operations. Feature fusion aims to use both the original feature information and hierarchical CLKAT feature information to enhance feature representation.

3.1.3 DFE

As shown in Fig.1(e), the DFE module mainly consists of deformable convolution [28]layers, 1×1 convolution layers, and ReLU activation functions. Traditional CNN models typically use convolution kernels with fixed and limited receptive fields, which cannot make full use of image feature information. Although the receptive field can be enlarged by expanding the convolution layers [4], this may cause grid artifacts in the generated images. Our model adopts DCN [28], which adaptively adjusts the shapes of the convolution kernels, enlarges the receptive field, and better captures relevant feature information in the image to improve model generalization. To obtain more detail information, two DCNs are concatenated to form the DFE module. DFE can dynamically change the resolution and receptive field size of feature maps based on the content and structure of the input image. This allows the network to better represent and preserve image detail information, enhancing the expressiveness and robustness of feature maps.

3.2 Loss Function

To effectively optimize the model, we designed a combination of multiple loss functions to update network parameters. The losses include the following:

Pixel Loss: This loss optimizes the network at the pixel level by minimizing the difference between the pixel values of the enhanced image and the reference image.

CR Loss [29]: This loss brings the information of the enhanced image closer to the reference image by contrasting the information between pairs of images.

Structural Loss: This loss optimizes the network based on SSIM to accurately restore underwater images.

The three losses are described in detail in the following sections, where denotes the distorted input image; denotes the reference image;$M(\cdot)$ denotes CAUIE; and the generated image is $M(x)$.

3.2.1 Pixel Loss

Pixel-wise loss is fundamental for image enhancement tasks. The L1 loss is a widely used loss function for single image restoration. We adopt the L1 loss as the pixel loss, which is calculated as follows:

$$L_1 = \frac{1}{w \times h} \sum_{i=1}^{h} \sum_{j=1}^{w} \left| M(x)(i,j) - y(i,j) \right| \tag{1}$$

Here, w and h denote the width and height of the generated image, respectively.

3.2.2 CR Loss

CR loss is a loss function proposed by Wu H et al. [29] for comparing overall information based on contrastive learning. CR loss constructs two sample pairs: one consists of the reference image and generated image and the other consists of the distorted image and enhanced image. The two pairs are forwarded through a VGG-19 [30] to

obtain two sets of features. The L1 losses of the two feature sets are calculated separately and weighted by coefficients for each layer to obtain the final CR loss. As an auxiliary loss function, CR can move the generated image closer to the reference image and farther from the distorted image. The CR loss is calculated as follows:

$$L_{contrastive} = \sum_{v=1}^{k} \sum_{i=1}^{h} \sum_{j=1}^{w} w_v \frac{M\left(x\left(i,j\right)\right) - y\left(i,j\right)}{M\left(x\left(i,j\right)\right) - x\left(i,j\right) + C} \tag{2}$$

In Equation (2),w_v is the weight for the v-th layer of the VGG-19, which was set to [1/32, 1/16, 1/8, 1/4, 1] in our experiments.C denotes a very small constant used to prevent division by zero errors.

3.2.3 Structural Loss

Structural loss is a loss function based on SSIM and can be used as a loss for image restoration tasks. To improve the network's ability to restore structural information in underwater image enhancement, the proposed method adopts SSIM as an optimization objective. The SSIM calculation is

$$SSIM\left(p\right) = \frac{\left(2u_\eta \cdot u_y + C_1\right) \cdot \left(2\delta_{\eta y} + C_2\right)}{\left(u_\eta^2 + u_y^2 + C_1\right) \cdot \left(\delta_\eta^2 \cdot \delta_y^2 + C_2\right)} \tag{3}$$

and the structural loss function is

$$L_{SSIM} = 1 - \frac{1}{N} \cdot \sum_{p \in P} SSIM\left(p\right) \tag{4}$$

where u_η and u_y denote the mean values of the enhanced and reference images, respectively,δ_η and δ_y denote the variances of the enhanced and reference images, respectively,$\delta_{\eta y}$ denotes the covariance,C_1 and C_2 denote constants, p denotes the pixels, and N denotes the number of pixels in image block P.

3.2.4 Mixed Loss

We use a weighted summation to calculate the total loss as the objective function for optimizing the model, as follows:

$$L_{total} = \alpha \cdot L_1 + \beta \cdot L_{contrastive} + \lambda \cdot L_{SSIM} \tag{5}$$

Here, the weights α,β, andλwere set to 0.16, 0.84, and 0.2, respectively, in the experiments.

4 Experiments and Analysis

To evaluate the performance of our model, comparative experiments were conducted on two public underwater image enhancement datasets: HICRD [7] and UIEBD [8].

HICRD[7] contains 2000 enhanced images and 6003 original images. UIEBD[8] has 950 real-world underwater images and the corresponding reference images for 890 of these images. These two datasets cover various underwater scenes and degradations, making them important for researching and improving underwater image enhancement algorithms.

Table 1 Hyperparameter settings for the experiments on the HICRD[7]

Hyperparameter	Value
Training image size	1842×980
Optimizer	AdamW
momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Learning rate	2e-4
Learning rate schedule	Cosine learning rate schedule
Batch size	2
Patch size	256×256
Epochs number	1000

Table 2 Results of the evaluation indicators for the HICRD

algorithm	PSNR(dB)	SSIM
UDCP[10]	13.31	0.493
DCP[9]	14.27	0.532
Haze-line[31]	14.69	0.423
IBLA [11]	19.42	0.463
UWCNN[8]	20.20	0.754
CycleGAN[32]	21.82	0.591
CUT[33]	26.30	0.796
CWR[15]	26.88	0.831
QACG[2]	<u>26.89</u>	<u>0.897</u>
Ours(CAUIE)	34.86	0.996

The red numbers in the table indicate the best results and the numbers underlined indicate the second best results

We used 1700 paired images in HICRD[7] as the training set and 300 images as the test set. The hyperparameters are listed in Table1. The experiments were conducted using Python 3.7 and PyTorch 1.13.1+cu116. Two evaluation metrics were used to compare the different methods: PSNR and SSIM. Both are full-reference metrics. PSNR measures the distortion of the images, whereas SSIM measures the similarity in structure. To evaluate the performance of CAUIE, we trained it for 1000 epochs on the HICRD[7], and compared it with other methods in terms of quantitative metrics and visual quality. The quantitative results on HICRD[7] are presented in Table 2, whereas

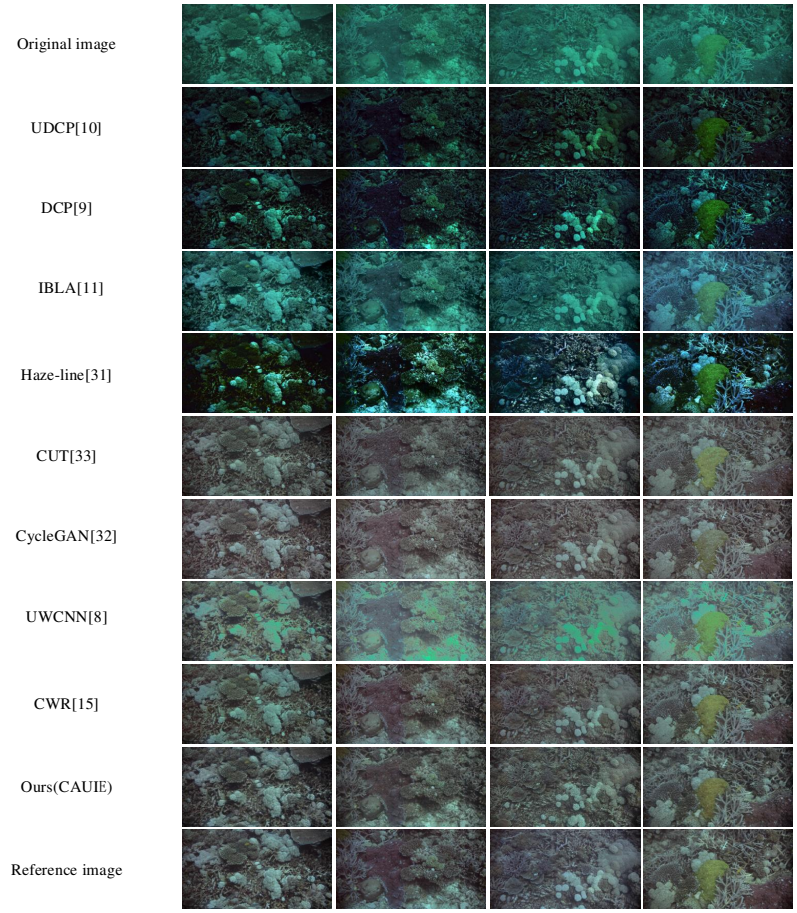

Fig. 2 Qualitative results obtained on the HICRD test dataset, where all examples were randomly selected from the test dataset. We compared our model with other underwater image restoration models. Traditional restoration methods fail to remove the green and blue color casts from the underwater images. Our model demonstrates satisfactory visual results without content and structural loss.

the subjective visual comparisons are presented in Fig. 2. From the results, it can be seen that on this dataset, CAUIE outperforms the previously best method QACG [2] by 7.97 dB in PSNR and 0.099 in SSIM. From a visual comparison with other methods, we conclude the following. DCP [9] and UDCP [10] are methods based on the DCP[9], which effectively removes haze in images but cannot handle color distortions. Haze-line [31] tends to lose detail or over-enhance the images, resulting in low quality and poor visual results in the enhanced images. IBLA [11] reduces haze and enhances the clarity of turbid underwater images, but excessive contrast stretching in some cases can lead to saturated bright and dark regions, causing detail loss and severe color distortions. UWCNN [17]cannot handle color distortions well, yielding results with a greenish appearance. Compared with other methods, CUT [33], CycleGAN [32], and

CWR [15] reduce color distortions better but may lose details, yielding blurry results. Our method achieves results closer to the reference images. In summary, the proposed algorithm obtains satisfactory underwater image enhancement performance, resolving color distortions while preserving image details.

Table 3 Results of the evaluation indicators for the UIEBD

algorithm	PSNR/dB	SSIM
Retinex[34]	17.53	0.773
IBLA[11]	18.51	0.762
WaterNet[8]	19.31	0.830
CUT [33]	20.34	0.765
USUIR [35]	20.31	0.841
CWR [15]	21.07	0.791
Fusion[36]	21.18	0.822
PUIENet[19]	<u>21.86</u>	<u>0.870</u>
Ours(CAUIE)	22.12	0.889

The red numbers in the table indicate the best results and the numbers underlined indicate the second best results

To further evaluate the effectiveness of CAUIE, we also trained it for 1000 epochs on the UIEBD[8]. Since the image sizes in UIEBD[8] are not unified, the batch size could only be set to 1 in this experiment. Table 3 shows a quantitative comparison of the different methods using the evaluation metrics. The results demonstrate that our method achieves a PSNR of 22.12dB and an SSIM of 0.889, which are slightly better than the results of the previously best method PUIENet [19].

A visual comparison of the results on UIEBD[8] is shown in Fig. 3. Fusion [36] can handle color distortions well, but may lose details or over-enhance the image during Fusion[36], resulting in unnatural visual results. IBLA [11] yields color distorted results with lost details. WaterNet [8] can adapt to different underwater images, but does not consider degradations like noise and artifacts in its design, which affects the quality and naturalness of the results and yields darker colors. PUIENet [19] captures underwater information more accurately and yields more natural visual results. CAUIE achieves results that are comparable to those of PUIENet [19] in terms of color correction and clarity, with good visual quality.

Additionally, ablation experiments were conducted to validate the impact of different loss combinations, as presented in Table 4. Seven groups with different loss functions were designed. The first group uses the L1 loss only, achieving the PSNR of 33.92 dB and the SSIM of 0.994. The second group uses SSIM loss only, achieving the PSNR of 33.13 dB and the SSIM of 0.992. The third group uses CR loss only, giving poor results, with the PSNR below 17, suggesting that CR loss alone is ineffective. The fourth group uses both L1 and SSIM losses with weights of 0.16 and 0.84, respectively, achieving the PSNR of 34.28 dB and the SSIM of 0.995. The fifth group uses SSIM and CR losses with weights of 1 and 0.2, respectively, achieving the PSNR of

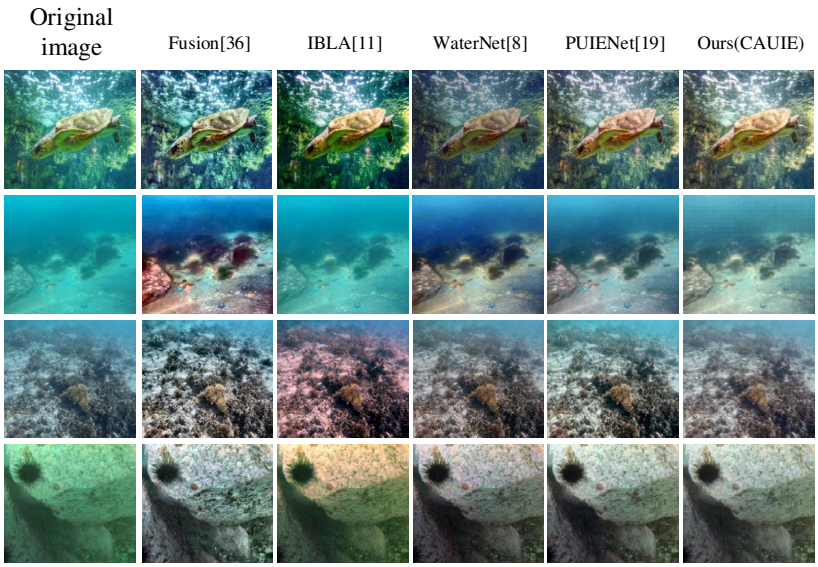

Fig. 3 Figure 3 Qualitative results on the UIEBD, where all examples were randomly selected from the test dataset. We compared our model with other underwater image restoration models. Traditional restoration methods fail to remove the green and blue color casts in underwater images. Among them, Ours and PUIENet [19] achieve the best performance, displaying superior visual results while minimizing content and structural losses.

Table 4 Loss ablation experiments on the HICRD

$L_1 loss$	SSIM loss	CR loss	PSNR/dB	SSIM
1	0	0	33.92	0.994
0	1	0	33.13	0.992
0	0	1	Nan	Nan
0.16	0.84	0	34.28	0.995
0	1	0.2	34.03	0.993
1	0	0.2	34.28	0.995
0.16	0.84	0.2	34.86	0.996

The values in the first three columns represent the weights of the loss function, respectively

34.03 dB and the SSIM of 0.993. The sixth group uses L1 and CR losses with weights of 1 and 0.2, respectively, achieving the PSNR of 34.28 dB and the SSIM of 0.995. The seventh group uses all three losses, L1, SSIM, and CR, with weights of 0.16, 0.84, and 0.2, respectively, achieving the highest results: the PSNR of 34.86 dB and the SSIM of 0.996.

A comparison of groups 1, 2, and 4 reveals that the mixed L1 and SSIM loss outperforms the use of either alone; the results of group 3 show that using only the

CR loss gives very poor results, with a PSNR that is less than 17. However, pairwise comparisons of groups 1 and 5, 2 and 6, and 4 and 7 show that adding the CR loss on top of the original loss improves the results and convergence speed. Therefore, an analysis of the experimental data reveals that using the CR loss alone does not benefit model optimization, but incorporating it as an auxiliary loss on top of existing losses improves results and convergence speed.

Table 5 Experimental comparison of ReLU and GELU

Structure	PSNR/dB	SSIM
CAUIE($ReLU$)	34.86	0.996
CAUIE($GELU$)	34.94	0.996

Song et al. [22] found that ReLU is more suitable than GELU for low-level computer vision tasks such as image dehazing. To validate this conclusion, we conduct a comparative experiment, as listed in Table 5. The first group is our proposed model using ReLU, and the second group uses the GELU in the attention [6]. The results confirm that using ReLU as the activation function yields better enhancement results than GELU, increasing the PSNR by 0.19 dB.

Table 6 Results of the CLKAT structural ablation experiments

CLKAT structure	Parameters	FLOPs	PSNR/dB	SSIM
Pre_IPLKA-CPLKA-Pos_IPLKA	2.8M+	126.02M+	34.78	0.996
Pre_IPLKA-CPLKA	1.9M+	91.52M+	30.90	0.993
CPLKA-Pos_IPLKA	1.9M+	91.52M+	34.86	0.996

The red numbers in the table represent the best results in that evaluation metric

The basic CAT[5] structure has three layers in sequence: IPSA-CPSA–IPSA. However, when compared with previous underwater image enhancement models, it has an excessive number of parameters. Our aim was to streamline it to obtain a model with fewer parameters and computations while retaining the enhancement performance. Therefore, we designed an ablation experiment with three groups: the first group uses the original three-layer structure, Pre_IPLKA–CPLKA–Pos_IPLKA; the second group removes Pos_IPLKA to obtain Pre_IPLKA–CPLKA; and the third group removes Pre_IPLKA to obtain CPLKA–Pos_IPLKA. The results are listed in Table 6.The loss function scheme used in this set of ablation experiments is the hybrid loss function with the best results described above, and its weights are shown in the last row of Table 4.

Table 7 reveals that removing Pos_IPLKA reduces the number of parameters by 0.9M but also decreases performance, reducing the PSNR to 30.90. In contrast, removing only Pre_IPLKA significantly reduces the number of parameters while improving

results, achieving a PSNR of 34.86. The experiments show that the original three-layer Pre_IPLKA–CPLKA–Pos_IPLKA structure is not optimal. On the bases of these results, to reduce the computations and parameters without sacrificing accuracy, we modified the original three-layer CAT[5] to the optimal two-layer structure CPLKA–Pos_IPLKA.

Table 7 Results of the DFE ablation study

DFE structure	Parameters	FLOPs	PSNR/dB	SSIM
w/o DFE	1.9M+	87.33M+	34.41	0.995
1-layer DCN	1.9M+	89.41M+	30.90	0.996
2-layer DCN	1.9M+	91.52M+	34.86	0.996

The red numbers in the table represent the best results in that evaluation metric

Additionally, we designed an ablation study with three groups to evaluate the effectiveness of DFE: the first group removes the DFE, the second group uses a single-layer DCN as the DFE, and the third group uses a two-layer DCN as the DFE in our model. The results are presented in Table 7. It can be seen that the three groups have negligible differences in the number of parameters and SSIM. The third group, which achieves the best performance, is our proposed model. The second group has a PSNR that is 0.61 dB lower than the PSNR of the third group. The first group performs slightly worse than the second group, but has a more noticeable 0.72 dB decrease with respect to the results of the third group.

Table 8 Comparison of the results of the different attention mechanisms on the HICRD

Attention	Dim	Parameters	PSNR/dB	SSIM
Self-Attention[26]	4	4.23M	31.65	0.990
Pool ($Pool\_size = 3$) [25]	4	11.20K	27.89	0.946
LKA ($Kernel\_size = 5$) [6]	4	14.75K	31.33	0.991

The red numbers in the table represent the best results in that evaluation metric

Given the excessive numbers of parameters and computations of the original CAT[5], we conducted an experiments to reduce them and compared the results obtained on HICRD[7], as presented in Tables 8. The loss function scheme used in this set of ablation experiments was the L_1 loss function only, and only 300 epochs of training were performed.One approach uses the pooling layers proposed in Poolformer [25] to replace the self-attention in the CAT. This significantly reduces the number of parameters from 4.23M to 11.20K, and the computation for one HICRD [7] image is reduced from 60G to 1G. However, the accuracy drops severely, from 31.65 dB to 27.89 dB. The other approach uses LKA[6] to replace self-attention in the CAT[5], as

proposed in this paper. This also greatly reduces the number of parameters to 14.75K and the computations to 1.3G per image, which is slightly more than the pooling approach, but this approach maintains excellent performance, with a PSNR of 31.33 dB, which is close to the original 31.65 dB. Therefore, it achieves the goal of reducing the parameters and computations while preserving image enhancement quality.

To maximize the performance of our model under the constraint of low computational requirements, we set the upper limit of computations to 10G and increased the hyperparameter dimensions as much as possible to 64. The loss function is the best performing one from previous experiments, and we increased the training epochs to 1000. This obtains the optimal results of our model on HICRD[7]: only 1.98M parameters, a PSNR of 34.86 dB, and an SSIM of 0.996. This is also the best result achieved in all our experiments.

5 Conclusion

This paper proposed a new underwater image enhancement model to address the color distortions and loss of details in underwater images. The model is based on the U-Net architecture and incorporates CAT, LKA, and DFE modules to tackle these problems. First, CAT effectively extracts global and local information from underwater images and fuses them, enabling better feature capture. Second, the modified LKA greatly reduces the numbers of parameters and computations of CAT while maintaining accuracy. This combination improves performance while reducing the computational burden. Additionally, the DFE module captures the detailed information of underwater images, enhancing contrast and color in the generated images for more clarity and realism. Moreover, the CR loss was introduced into the hybrid optimization objective. This loss enables consistency between the enhanced image and the ground truth with respect to content, details, and color. The proposed model was evaluated on two public underwater image datasets using subjective and objective comparisons. The results demonstrate the superiority of our proposed algorithm over the comparison methods, proving its effectiveness for underwater image enhancement.

6 Acknowledgment

This work was supported by National Natural Science Foundation of China (No.62172045 and No.62272049).

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

[2] Lan, Z., Zhou, B., Zhao, W., Wang, S.: An optimized gan method based on the que-attn and contrastive learning for underwater image enhancement. Plos one **18**(1), 0279945 (2023)

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[4] Ronneberger, O., Fischer, P., Brox, T.: 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, pp. 234–241 (2015). Springer

[5] Lin, H., Cheng, X., Wu, X., Shen, D.: Cat: Cross attention in vision transformer. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2022). IEEE

[6] Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., Hu, S.-M.: Visual attention network. Computational Visual Media, 1–20 (2023)

[7] Han, J., Shoeiby, M., Malthus, T., Botha, E., Anstee, J., Anwar, S., Wei, R., Armin, M.A., Li, H., Petersson, L.: Underwater image restoration via contrastive learning and a real-world dataset. Remote Sensing **14**(17), 4297 (2022)

[8] 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**, 4376–4389 (2019)

[9] He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010)

[10] Drews, P., Nascimento, E., Moraes, F., Botelho, S., Campos, M.: Transmission estimation in underwater single images. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 825–830 (2013)

[11] Peng, Y.-T., Cosman, P.C.: Underwater image restoration based on image blurriness and light absorption. IEEE transactions on image processing **26**(4), 1579–1594 (2017)

[12] Akkaynak, D., Treibitz, T.: A revised underwater image formation model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6723–6732 (2018)

[13] Akkaynak, D., Treibitz, T.: Sea-thru: A method for removing water from underwater images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1682–1691 (2019)

[14] Cao, K., Peng, Y.-T., Cosman, P.C.: Underwater image restoration using deep networks to estimate background light and scene depth. In: 2018 IEEE Southwest

Symposium on Image Analysis and Interpretation (SSIAI), pp. 1–4 (2018). IEEE

[15] Han, J., Shoeiby, M., Malthus, T., Botha, E., Anstee, J., Anwar, S., Wei, R., Petersson, L., Armin, M.A.: Single underwater image restoration by contrastive learning. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 2385–2388 (2021). IEEE

[16] Li, J., Skinner, K.A., Eustice, R.M., Johnson-Roberson, M.: Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. IEEE Robotics and Automation letters **3**(1), 387–394 (2017)

[17] Li, C., Anwar, S., Porikli, F.: Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recognition **98**, 107038 (2020)

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

[19] Fu, Z., Wang, W., Huang, Y., Ding, X., Ma, K.-K.: Uncertainty inspired underwater image enhancement. In: European Conference on Computer Vision, pp. 465–482 (2022). Springer

[20] Kar, A., Dhara, S.K., Sen, D., Biswas, P.K.: Zero-shot single image restoration through controlled perturbation of koschmieder's model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16205–16215 (2021)

[21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[22] Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Transactions on Image Processing **32**, 1927–1941 (2023)

[23] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)

[24] Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)

[25] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10819–10829 (2022)

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

[27] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[28] Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)

[29] Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10551–10560 (2021)

[30] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[31] 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), 2822–2837 (2020)

[32] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

[33] Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, pp. 319–345 (2020). Springer

[34] Zhang, S., Wang, T., Dong, J., Yu, H.: Underwater image enhancement via extended multi-scale retinex. Neurocomputing **245**, 1–9 (2017)

[35] Fu, Z., Lin, H., Yang, Y., Chai, S., Sun, L., Huang, Y., Ding, X.: Unsupervised underwater image restoration: From a homology perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 643–651 (2022)

[36] Ancuti, C., Ancuti, C.O., Haber, T., Bekaert, P.: Enhancing underwater images and videos by fusion. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 81–88 (2012). IEEE