Preprints.org

Article

Not peer-reviewed version

Enhancement of Underwater Images through Parallel Fusion of Transformer and CNN

Xiangyong Liu , Zhixin Chen , Zhiqiang Xu * , Ziwei Zheng , Fengshuang Ma , Yunjie Wang

Posted Date: 19 July 2024

doi: 10.20944/preprints202407.1575.v1

Keywords: image enhancement; local features; global features; parallel fusion

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Enhancement of Underwater Images through Parallel Fusion of Transformer and CNN

Xiangyong Liu [1,2], Zhixin Chen [1], Zhiqiang Xu [1,*], Ziwei Zheng [3], Fengshuang Ma [1] and Yunjie Wang [1]

[1] Fishery Machinery and Instrument Research Institute, Chinese Academy of Fishery Science, Shanghai, 200092, China.; liuxiangyong@fmiri.ac.cn (X.L.); chenzhixin@fmiri.ac.cn (Z.C.); mafengshuang@fmiri.ac.cn (F.M.); wangyunjie@fmiri.ac.cn(Y.W.)
[2] State Key Laboratory of the Internet of Things for smart city (IOTSC), University of Macau, 999078, Macau.
[3] Digital Industry Research Institute, Zhejiang Wanli University, No.8 South Qian Hu Road, Ningbo City, Zhejiang Province, 315199, China; zhengziwei@tsinghua.org.cn (Z.Z.);
* Correspondence: xuzhiqiang@fmiri.ac.cn

Abstract: Ocean exploration is crucial for utilizing its extensive resources. Images captured by underwater robots suffer from issues such as color distortion and reduced contrast. To address the issue, we propose an innovative enhancement algorithm that integrates Transformer and Convolutional Neural Network (CNN) in a parallel fusion manner. Firstly, a novel transformer model is introduced to capture local features, employing peak-signal-to-noise ratio (PSNR) attention and linear operations. Subsequently, to extract global features, both temporal and frequency domain features are incorporated to construct convolutional neural network. Finally, the Fourier's high and low-frequency information of the original image are utilized to fuse different features. To demonstrate the algorithm's effectiveness, underwater images with various levels of color distortion are selected for both qualitative and quantitative analyses. The experimental results demonstrate that our approach surpasses other mainstream methods, achieving superior PSNR and structural similarity index measure (SSIM) metrics and leading to a detection performance improvement of over ten percent.

Keywords: image enhancement; local features; global features; parallel fusion

1. Introduction

Exploration of the ocean is vital for harnessing its abundant resources [1]. Underwater robots are crucial instruments to explore the ocean, which enables image-based target detection tasks. Due to light attenuation and scattering in seawater, the quality of these images is often compromised. Consequently, underwater image enhancement algorithms are critical in correcting these distortions [2] (Figure 1), establishing a significant research area within the fields of computer vision and underwater robotics. Four methods [3] are selected to test the same image with our computer, showcasing different repaired quality. Therefore, underwater image processing faces significant challenges due to color distortion and reduced contrast caused by the absorption and scattering effects of water.

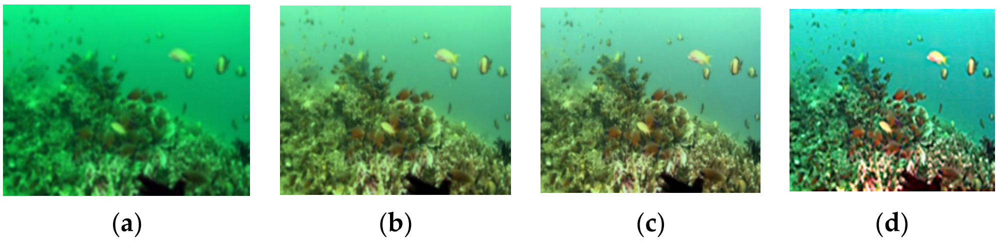


Figure 1. Visual comparisons on a real underwater image. Different methods appear different color deviation and image resolution [3]. (a) MIR-Net. (b) U-net. (c) WaterNet. (d) Ucolor.

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

Underwater Research in image enhancement focuses on improving the quality of distorted underwater images, with a focus on restoring color distortion information [4]. Some scholars have explored non-deep learning approaches and have made some progress. Non-deep learning methods rely on statistical assumptions and model to enhance underwater images, such as Underwater Dark Channel Prior (UDCP) [5], Image Blur Recovery [6], and Underwater Light Attenuation Prior (ULAP) [7]. Cheng et al. [8] pointed out that the dissolved substances in water can weaken the imaging process, and influence the attenuation parameters of light propagation in water. Drews et al. proposed an underwater prior method by utilizing red channel information [9]. Li et al. proposed an underwater light attenuation prior (ULAP) model to restore image quality [10]. Ma et al. devised a wavelet transform network that decomposes input images into frequency maps to enhance image details [11]. However, the complexity of underwater environments often leads to inaccuracies in parameter estimation for these methods.

Currently, neural networks have been widely employed to various visual tasks [12]. In contrast, extensive datasets and specialized loss functions have been utilized by deep learning techniques to train deep neural networks for image quality enhancement, including models like Underwater Residual Network (UResNet), Shallow Underwater Network (UWNet), and Underwater Convolutional Neural Network (UWCNN) [13]. Mean square error loss and edge difference loss are used to optimize convolutional neural networks for image enhancement [14]. By employing conventional convolutions, Naik et al. developed a network specifically for underwater image enhancement, which demonstrates effective enhancement capabilities on public datasets [15]. Li et al. introduced a residual network-based underwater image enhancement algorithm [16]. Chen et al. introduced an end-to-end neural network enhancement model that integrates residual structures and attention mechanisms [17]. Current enhancement algorithms predominantly rely on convolutional neural networks, but they often utilize a single feature extraction backbone. However, the features extracted from these models are often insufficiently detailed.

Wang et al. utilized Generative Adversarial Networks (GANs) to design a feature enhancement network [18,19]. Moreover, the underwater generative adversarial net-works (UGANs) [20] scheme has been established for UIE task by using encoder–decoder structure [21,22], whereby the preservation of rich semantic information can be achieved. Junjun Wu et al. have developed a multi-scale fusion generative adversarial network named Fusion Water-GAN (FW-GAN), which aims to improve underwater image quality while effectively preserving rich semantic information. This network integrates four convolutional branches to achieve this goal [23]. Kei et al. created a dataset that includes both image and sonar data specifically designed for low-light underwater environments, utilizing a Generative Adversarial Network (GAN) to improve image quality. Experimental results show that this method achieves better detection performance [24]. Zhang et al. collected images from different angles and then calculated the camera poses for each angle. They fed the collected image sequences and their corresponding poses into a Neural Radiance Field (NeRF), synthesizing new viewpoints and improving the effect of 3D image reconstruction [25]. Adversarial learning methods are mostly based on object detection with similar quality or visibility, and acquiring clear sample data for these models remains a formidable task.

Deep learning encompasses various backbone architectures, including the widely utilized Convolutional Neural Networks (CNNs) and Transformers [26,27], which has gained popularity in computer vision tasks. The Transformer architecture incorporates features like multi-head mechanisms and multi-layer perceptions, making it versatile for a range of visual tasks [28]. Zamir et al. [29] employed an encoder-decoder structure to obtain features at different scales, achieving image enhancement in rainy and foggy weather conditions. Song et al. modified attention modules within the network layers, constructing a parameter-adjustable dehazing network [30]. Although the Transformer architecture shows great potentiality for computer vision tasks, its high computational complexity often results in increased computational load and longer processing times.

Despite advancements in current methods for addressing underwater image distortions, challenges persist in achieving high-quality restoration. To produce high-quality images, our research try to overcome the uncertainty in generated images by utilizing the complementary

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

features of different neural network frameworks. This paper introduces a new model for enhancing the quality of underwater images, aiming to tackle issues such as color shifts, unrealistic colors, and reduced contrast [31]. Furthermore, a matrix linear computation approach has been designed to minimize the computational delays caused by network stacking. To this point, an innovative approach is proposed to extract both local and global image features. This network integrates visual Transformer models with CNN networks to enhance the overall restoration process. Additionally, information fusion weights are calculated from the Fourier transform features of original image. The main advantage of our work are as followed:

1) A novel Transformer model that extracts local features has been proposed. It incorporates PSNR attention and linear operations to significantly reduce computational load and alleviate color artifacts.

2) Additionally, a novel global feature extraction network is devised, which leverages both temporal and frequency domain characters to enrich image features.

3). Additionally, a feature fusion method, utilizing the Fourier transform of the original image, has been introduced to optimize global feature weights through high-frequency Fourier transforms and local feature weights via low-frequency Fourier transforms.

2. Materials and Methods

Figure 2 depicts our detection framework. This network incorporates both CNN and Transformer backbones, which is designed to extract both global and local features, respectively. These extracted features are fused at the smallest down-sampling size. Additionally, the low-frequency and high-frequency information of the original image is obtained via Fourier transform, serving as fusion weights for the extracted CNN and Transformer features.

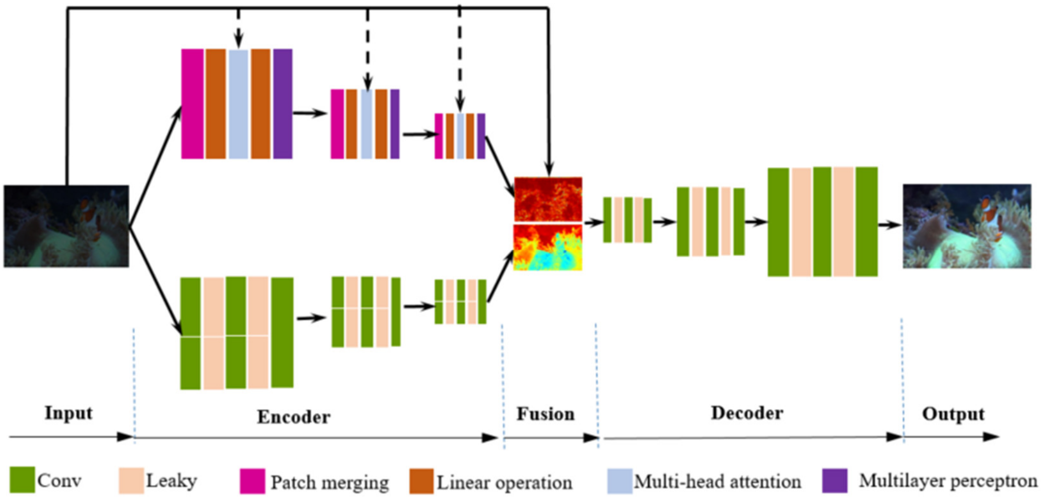


Figure 2. Architecture of the proposed network, which consist of three parts: encoder, fusion, and decoder.

2.1. Two branches' Feature Extraction Network

Conventional low-light image enhancement networks typically employ convolutional structures within the feature layers, predominantly extracting image information from the image's bright regions. These regions are rich in visible content or signals. However, in some non-prominent regions, fine-grained features may be lost, leading to a decrease in detection accuracy.

To this point, we employ two backbones for image enhancement. Information from both global and local regions complements each other, and this differentiation can be determined based on the distribution of Fourier transforms on the image. On one hand, for image regions characterized by high-frequency Fourier transforms, their features predominantly manifest in the globally salient information. On the other hand, in image regions corresponding to low-frequency Fourier

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

transforms, the Transformer predominantly captures local detailed information. This methodology facilitates the extraction of diverse image features.

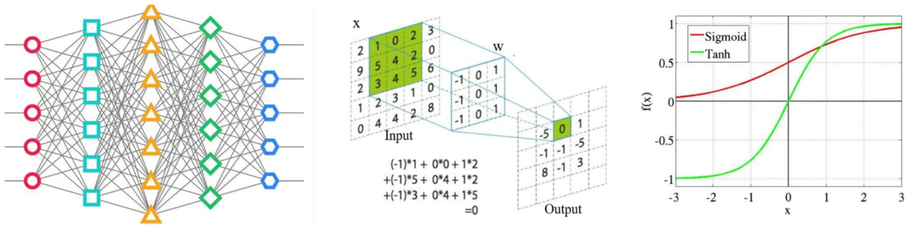
As to the encoder framework, we devised two different backbones for feature extraction. One employs transformers, while the other utilizes convolutional structures. Each backbone incorporates a top-down feature pyramid extraction network, segmented into three levels. The input dimension of the first layer features is 6 dimensions, yielding 64 dimensions as output. Both the second and third levels' feature extraction operations utilize 64 channels.

CNNs excel at extracting edges, textures, and simple shapes from images. Transformers, on the other hand, excel at identifying long-range dependencies and inferring local information. By integrating the local features extracted by Transformers with the global textures identified by CNNs, we can produce richer and more diverse representations. This combined approach is more effective at handling noise and variations in data. Moreover, hybrid models can better adapt to different types of data, which can deal with spatial and sequential information simultaneously. Thereby, the method can enhances the model's recognition and classification capabilities.

2.2. Implementation of the CNN Branch

Convolutional Neural Networks (CNNs) is a types of deep learning model specifically designed to process image data. CNNs employ convolutional layers for extracting local features, pooling layers to diminish data dimensionality and computational complexity, and fully connected layers for classification. The core advantages of CNNs lie in their local connections and shared weight, which make them particularly effective for image recognition and classification tasks.

Convolutional Neural Networks extract features by sliding convolutional kernels over the pixels' matrix of the input image, computing the weighted sum of local regions to generate feature maps (Figure 3 (b)). The kernels capture local features like edges and textures. Subsequently, activation functions perform nonlinear mappings on these features, enhancing the model's ability to filter key characteristics. The parameters of the network are optimized through the error back propagation and iterative learning. This learning process automatically constrains the input and maximizes the activation of the output.



(**a**) (**b**) (**c**)

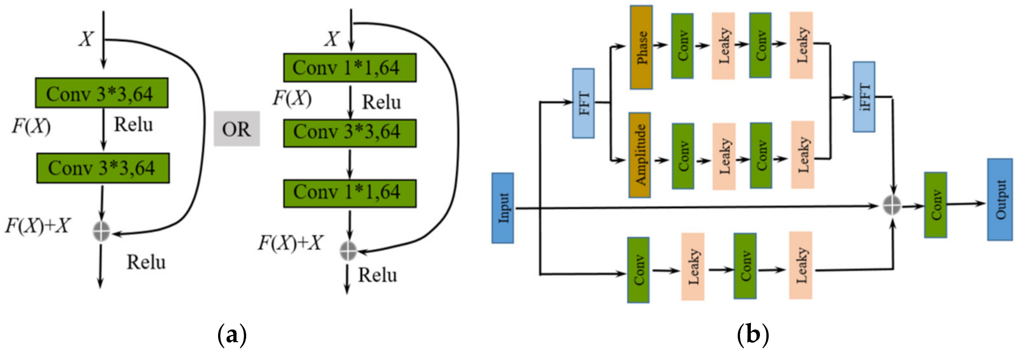
Figure 3. The learning process for the optimal model [32]. (a) Convolutional Neural Network. (b) The convolutional kernels. (c) The activation functions.

Residual Network (ResNet) is an improved version of CNNs, introducing skip connections or residual connections (Figure 4 (a)). These connections allow gradients to pass directly from later layers to earlier layers, addressing the issues of vanishing and exploding gradients in training deep networks. By stacking more layers, ResNet achieves deeper network architectures than the traditional CNNs, and have demonstrated significant performance in various visual tasks [33].

Different from the ResNet network, some researchers have noted that the brightness degradation on images primarily resides in the magnitude components of Fourier transformation, while the rest exists in the phase components [34]. Inspired by previous research on Fourier transformation, this backbone further introduces the correlation properties between magnitude component and brightness to enhance feature extraction effectiveness. In this backbone, we designed two stages' feature architecture (Figure 4 (b)). In stage one, brightness of low-light image features is enhanced by

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

optimizing the amplitude in Fourier space. In stage two, features from convolutional neural networks are further integrated.



(a) (b)

Figure 4. Comparisons for different CNN feature-extraction network unite. (a)The commonly used ResNet network. (b) Our feature network incorporates both temporal and frequency domain characters.

In stage one, given an input image *x* with a dimensions of *H*×*W*, its transformation into the frequency space is represented by Equation (1):

$$F(x)(u,v)=X(u,v)=\frac{1}{\sqrt{HW}}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}x(h,w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)} \tag{1}$$

Where, *h* and w denote coordinates in the temporal domain, while *u* and *v* represent coordinates in the frequency domain. To extract the amplitude and phase components, the Fourier processing (FFT) block extracts frequency features (Figure 4). Subsequently, two 1x1 convolutional layers with Leaky activation are applied to each branch. Finally, an inverse Fourier transform (iFFT) is applied to convert these two branches back to the spatial domain.

The Fourier transformation primarily relies on convolutional changes in the frequency domain to enhance brightness, while lacking convolution operations in the temporal domain for extracting details. Therefore, in the second stage, convolution operations in the temporal domain are employed to enrich features. Finally, to achieve the ultimate feature fusion, dimension addition and reduction operations are separately applied to the features in the temporal-frequency domain.

Compared with the commonly used residual network (Figure 4 (a)), our method with the Fourier Transform(Figure 4 (b)) can convert convolution operations into multiplication operations in the frequency domain, thereby enhancing the efficiency of convolution calculations. By enhancing specific frequency components, specific edge features can be accentuated, which is especially useful for image enhancement. However, during the inverse Fourier Transform process, some important information may be lost. Therefore, integrating features from the temporal domain is crucial to preserve specific characteristics.

2.3. Implementation of the Transformer Branch

Unlike CNNs, which extract features through global attentions, transformers capture local features from token regions. Local feature extraction is achieved through the matrixes' multiplication, which has been validated in various high-level and low-level tasks. Assuming the feature dimensions are *h***w***C* and the token size is *p***p*, the total number of feature tokens can be calculated as *m*=(*h*/*p*)*(*w*/*p*)**C*. Moreover, multi-head self-attention (MSA) modules and multi-layer perceptions (MLP) are employed in transformer. Assuming the input features to the transformer have the same dimensions, tokens are merged into a sequence of features with multiple heads (Figure 5 (a)). $Q*K^T$=(R^{d*n})*(R^{d*n})^T= R^{d*d}, and its computation load= d^2*n. Then, (R^{d*d})*(R^{d*n}) = R^{d*n}, and its computation load= d^2*n. In summary, the total computation load= 2d^2*n. The calculation process of feature transformation and computational complexity are respectively depicted in Equation (2): and Equation (3):

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

$$\begin{cases} q_i = k_i = v_i = LN\left(F_1, F_2, \ldots, F_m\right) \\ \hat{y}_i = MSA\left(q_i, k_i, v_i\right) + y_{i-1} \\ y_i = MLP\left(LN\left(\hat{y}_i\right)\right) + \hat{y}_i \end{cases} \tag{2}$$

$$\hat{x} = \text{softmax}\left(\frac{Q * K^T}{\sqrt{d}}\right) * V \tag{3}$$

Where, Q is the query matrix, Q is the key matrix, and Q is the value matrix.

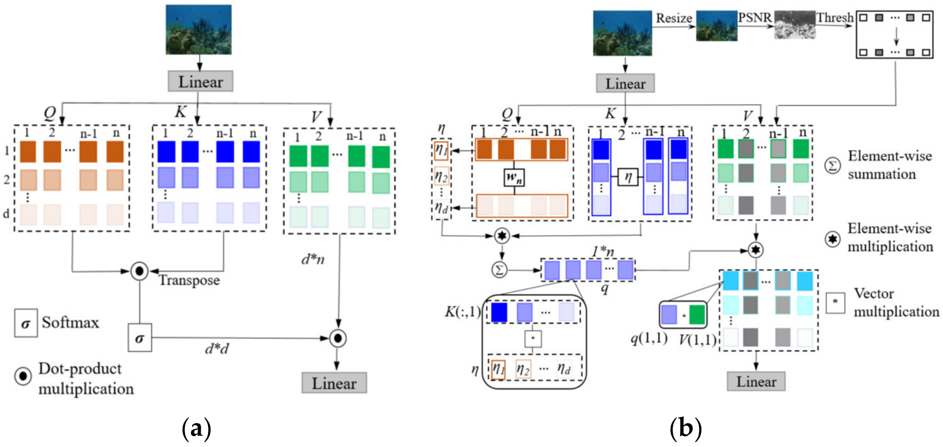


Figure 5. The transformer feature-extraction network with PSNR-based attention and linear operations. (a) The dot-product multiplication operation. (b) The element-wise multiplication operation with PSNR-based attention.

To alleviate calculation complexity, the method in Figure 5 (b) replaces traditional dot-product multiplication with element-wise multiplication. Typically, *Q*, *K*, and *V* each have dimensions of $R^{d\times n}$. To compute the attention weights for extracting features from underwater images, the query matrix is initially multiplied by a trainable parameter vector ($w_n \in R^n$). This process results in the generation of a global attention vector of size η_d in Equation (4):

$$\eta_d = \frac{\exp\left(Q * w_n / \sqrt{n}\right)}{\sum_{j=1}^{d} \exp\left(Q * w_n / \sqrt{n}\right)} \tag{4}$$

Next, the *K* matrix undergoes element-wise multiplication with the global attention vector η_d to yield the global query vector *q*. As shown in Figure 5.(b), $q \in R^{1*n}$. Subsequently, this global vector *q* is element-wise multiplied with the *V* matrix to generate global features that merge information from both the *Q*-matrix and *K*-matrix. Unlike previous dot-product computations, the computational load of element-wise multiplication is linearly related to the parameters (*d*n*), alleviating overall computational load. Following this, we perform another transformation to activate the final information in Equation (5):

$$\begin{cases} q = \sum_{i=1}^{d} \eta_i * K_i \\ x = T\left(q * V\right) \end{cases} \tag{5}$$

Where *T* denotes the activation operation. To mitigate the influence of extremely dark regions on inference, SNR map is utilize to guide the learning attention of the transformer. For an input image $I \in R^{H\ x\ W\ x\ 3}$, with its corresponding SNR map $S \in R^{H\ x\ W}$, *S* is adjusted into $S' \in R^{h\ x\ w}$ to align with the dimensions of the feature map *F*. Then, *S′* is partitioned into *m* patches, and the average value for each patch is calculated. $S_i \in [0,1]$, where $i=\{1, \ldots, m\}$. This masking mechanism effectively prevents the influence of features with very low signal-to-noise ratio (SNR), as illustrated in Figure 5(b). The the *i*-th mask value of *S′* is designed in Equation (6):

$$S_i = \begin{cases} 0, & S_i < s \\ 1, & S_i > s \end{cases}, i = \{1, \ldots, m\} \tag{6}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

The masking calculation process for the x parameter is expressed in Equation (7):

$$\hat{x}=x*S \tag{7}$$

As shown in Table 1, the total calculation load is $3d*n$, which is far less than the previous calculation load $2*d^2*n$. According to the commonly used dot-product multiplication in Figure 5 (a), the computational complexity of the self-attention mechanism scales quadratically with the sequence length (d^2), causing a significant increase in resource consumption when the sequence length is large. Due to the large number of parameters in each layer, Transformer models are typically much larger than CNNs, requiring more memory and computational power for training and inference. By adopting the proposed hybridized block modular approach, the computational load can be reduced from $2d^2n$ to $3dn$, offering a substantial advantage. The integration of CNN and Transformer models increases complexity and computation time. Therefore, we reduce the algorithm's complexity through matrixes' element-wise multiplication.

Table 1. Computation load for different parameters**.**

Parameters	Computation load	Computation load summation
η_d	$d*n$	$3d*n$
q	$d*n$	
$\hat{x}$	$d*n$	

It is worth noting that similar block or token attention has been developed. Zhou et al. divided large images into smaller blocks, utilizing a trained rCNN as a block descriptor for image forgery detection [35]. Bei et al. introduced the block matching and grouping criterion, applying a convolutional neural network (CNN) within each block for 3D filtering to develop a well-suited denoising model [36]. Abbas et al. created an innovative hybrid block-based neural network model, integrating expert modular structures and divide-and-conquer strategies with a genetic algorithm (GA) [37]. To generate high-resolution landslide susceptibility maps, each sub-network module employs input blocks, layers of hidden blocks, and an additional decision block (Figure 6 (a)). Different from the independent block's method, the element-wise multiplication operation is developed in the research to extract the cross-attention (Figure 6 (b)).

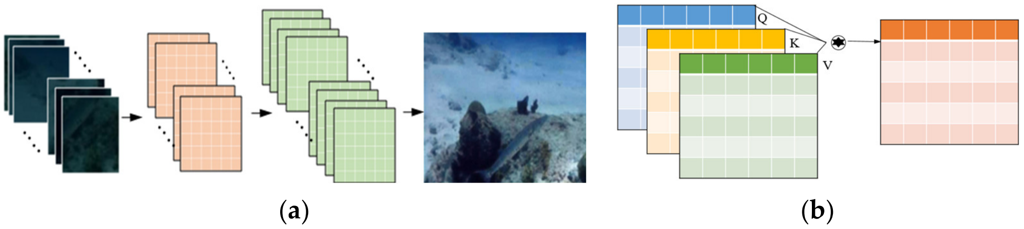
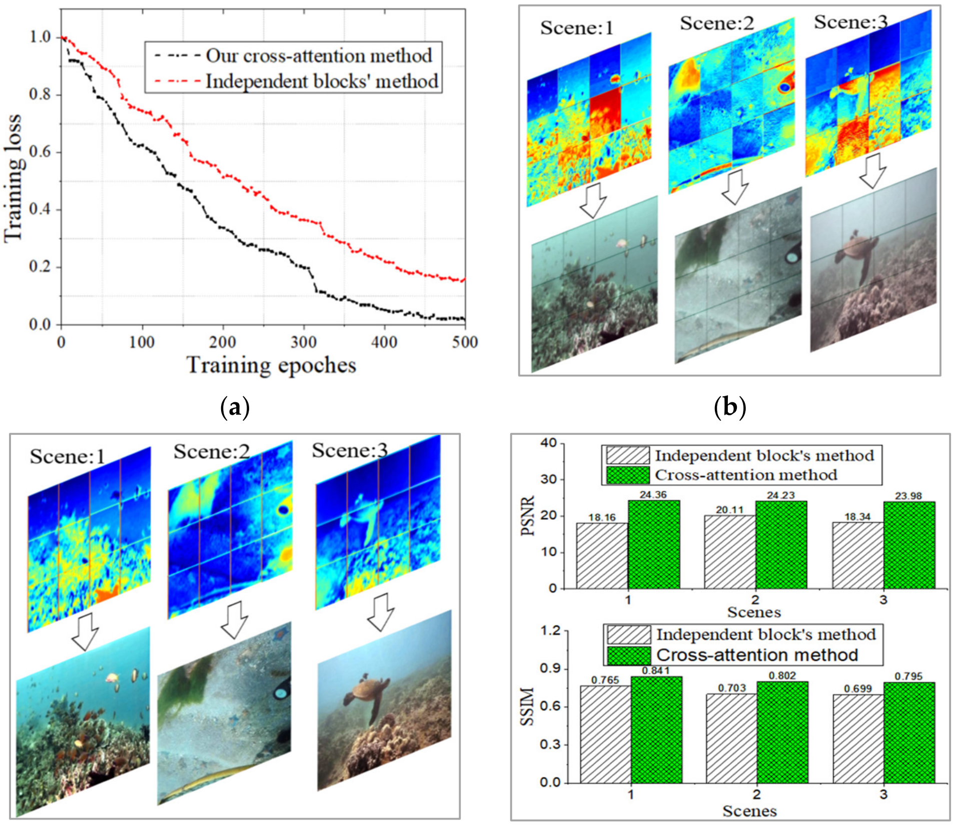


Figure 6. Comparison of different image block-based or modular structures. (a) The independent block's method. (b) The proposed cross-attention method with element-wise multiplication operation.

In order to compare the performance of different block based methods, the two structures in Figure 6 were used to train and enhance images, respectively. Images from three different scenarios are displayed. Despite the significant development and approved capability of image processing systems through the advanced block-based or modular structures, our presented model in this study offers three significant advantages. Firstly, it can reduce learning losses and accelerate model convergence (Figure 7 (a)). Secondly, it captures global forward-backward attention more effectively, and extracts the continuous features, which reduces information loss caused by independent blocks (Figure 7 (b) and (c)). Thirdly, it brings high-quality enhanced images with higher PSNR and SSIM indexes (Figure 7 (d)).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1



(a) (b)

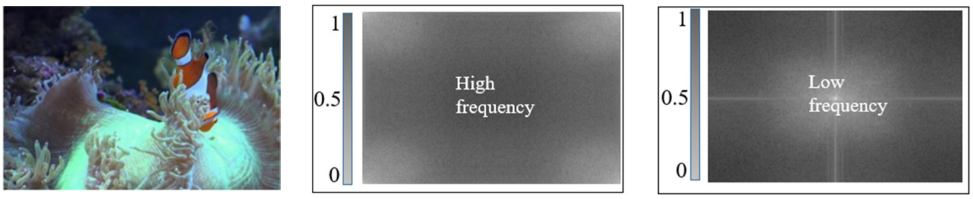
(c) (d)

Figure 7. Experimental verification for different image block-based or modular structures. (a) The training loss. (b) The learned discontinuous features of independent block's method. (b) The learned successive features of cross-attention method. (d) The PSNR and SSIM performance for different enhanced images.

2.4. Fusion Attention Based on High-Pass and Low-Pass Filters

The torch.add method is commonly used to perform element-wise addition of tensors [35]. This method checks the shapes of the input tensors. When the shapes are aligned, the addition operation is performed element by element. This means that corresponding elements from each tensor are added together, producing a new tensor as the result. The result of the addition operation can be stored into a new tensor. While this method can combine different features, it cannot differentiate or utilize the advantages of different features.

Different from the traditional torch.add method, the significant difference of low-frequency and high-frequency features is valuable and can be utilized. The display of an image relies on trigonometric frequency components. High-frequency signals cause rapid changes, leading to sharp edges within the image. Conversely, low-frequency signals induce more gradual changes, contributing to smoother appearance within the image. The role of filters is to pass or suppress certain frequency components of an image.



(a) (b) (c)

Figure 8. Fourier transformation of the image. (a) Original image. (b) Fourier transformation. (c) Shifted frequency.

The Fourier transform serves as a bridge between the temporal domain and the frequency domain (Figure 6). Ideal low-pass filtering, a method for image smoothing, retains low-frequency components. The transfer function of an ideal low-pass filter is represented in Equation (8):

$$\begin{cases} D(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-j2\pi(ux/M+vy/N)} \\ L(u,v)=e^{-D^2(u,v)/2D_0^2} \\ L(x,y)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} L(u,v)e^{j2\pi(ux/M+vy/N)} \end{cases} \quad (8)$$

Where M and N denote the length and height of the image, respectively. $D(u, v)$ denotes the frequency domain of the image, and $f(x, y)$ represents the temporal domain of image. The range of u is [0, M−1], and the range of v is [0, N−1]. $D(u, v)$ denotes the distance from the point (u, v) in the frequency domain to the center, while D_0 denotes the cutoff frequency. $L(u, v)$ denotes the low-pass filters in the frequency domain. $L(x, y)$ denotes the low-pass results in the temporal domain. Figure 9 illustrates the corresponding low-pass filter functions and their corresponding filtering outcomes.

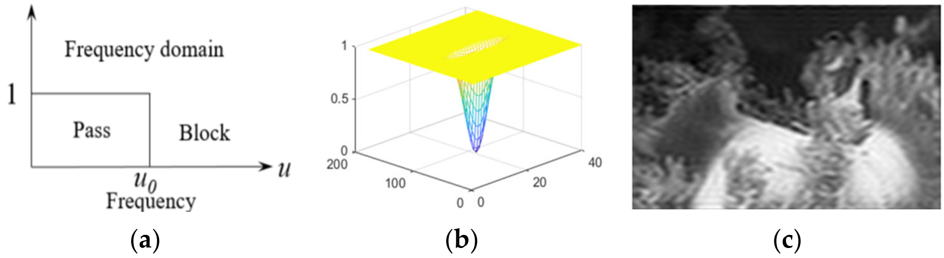


Figure 9. Low-pass filter. (a) Ideal low-pass filter. (b) Gaussian low-pass filter. (c) Low-pass filter result.

Different from the low-pass filters, high pass filters enhance the details and edges of an image by removing low-frequency components. The basic principle is to set the low-frequency components in the frequency domain to zero and only retain the high-frequency components. $H(u, v)$ denotes the high-pass filters in the frequency domain. $H(x, y)$ denotes the high-pass results in the temporal domain. Figure 10 illustrates the associated high-pass filter functions and their filtering results. The transfer function of a high-pass filter is represented in Equation (9):

$$\begin{cases} D(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-j2\pi(ux/M+vy/N)} \\ H(u,v)=1-e^{-D^2(u,v)/2D_0^2} \\ H(x,y)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} H(u,v)e^{j2\pi(ux/M+vy/N)} \end{cases} \quad (9)$$

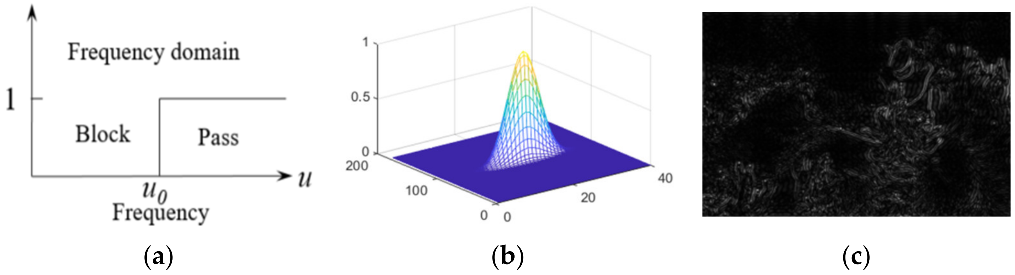


Figure 10. High-pass filter. (a) Ideal high-pass filter. (b) Gaussian high-pass filter. (c) High-pass filter result.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

We utilize Fourier transforms to calculate fusion weights for integrating different backbone features (Figure 11 (b)). Low-frequency features are extracted by the Transformer to align with locally smooth features. Sharp edge details are captured by convolutional neural networks to match the high-frequency features. Both high and low-frequency features are normalized to the [0-1] range. The different fusion calculation is illustrated in Equation (10):

$$\begin{cases} F^{'} = F_{CNN} + F_{Transformer} \\ F^{''} = F_{CNN} * H\left(x, y\right) + F_{Transformer} * L\left(x, y\right) \end{cases} \tag{10}$$

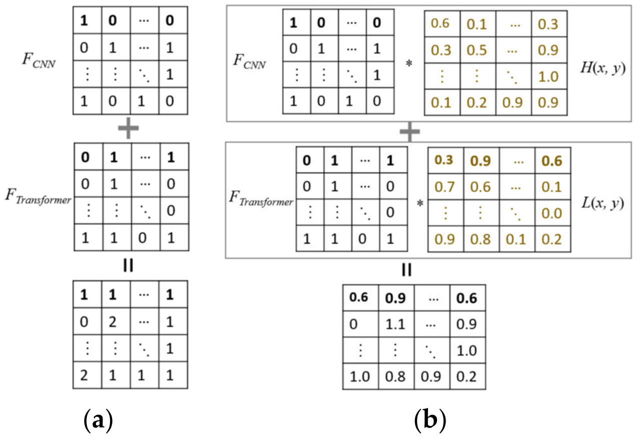


Figure 11. Comparison for different feature fusion method. (a) The commonly used torch.add method add the CNN and Transformer features. (b) The proposed approach incorporates different features with the weights of low-frequency and high-frequency features.

Low-frequency Fourier variations correspond to Transformer features, while high-frequency Fourier variations correspond to CNN features. Compared with the commonly used torch.add in Figure 11 (a), the combination of CNNs and Transformers can fully leverage the different advantages. This method not only enhances feature representation capabilities and robustness, but also optimizes the use of computational resources.

3.0. Experimental Validation

3.1. Dataset and Experimental Designation

We evaluate the algorithm's performance by using two publicly accessible datasets: LSUI [36] (Large-Scale Underwater Image dataset) and UIEB dataset [37]. LSUI comprises 5000 underwater images with varying exposure levels. The UIEB dataset includes pairs of low-exposure and high-exposure images, with 800 pairs designated for training, 150 pairs for validation, and 90 pairs for testing. The LSUI and UIEB datasets play crucial roles in underwater image enhancement research. LSUI, with its large and diverse data volume, offers ample material for training and testing deep learning models. Due to its high-quality annotated image pairs, UIEB is a key resource for evaluating and optimizing algorithms. By utilizing the two datasets, it provide us more powerful and robust underwater image enhancement algorithms, offering higher quality image processing solutions.

Our framework was implemented in PyTorch [38], and the training and testing processes are conducted on a computer equipped with a 2080Ti GPU. Gaussian distribution was used to randomly initialize the network training parameters. And standard data augmentation techniques, such as vertical and horizontal flipping, are applied. Our encoder frame includes three layers, which are followed by a feature fusion module. Similarly, the decoder comprises three layers, utilizing ChannelShuffle for up-sampling operations. Adam optimizer [39] with an initial learning rate of 1*e*-3 was used to minimize loss. The learning rate was decreased by 0.1 after every 100 iterations.

During training, we evaluated the model's performance through loss functions (such as MSE, PSNR, etc.), which measures the discrepancy between the output and ground truth images. The model's weights are saved during each epoch. The .ckpt files are used to save training weights. The loss function is expressed in Equation (11):

$$\text{Total Loss}=\alpha^{*}\text{MSE}+\beta^{*}(1-\text{SSIM})+\gamma^{*}\text{PSNR} \tag{11}$$

Here, α, β, and γ are the weighting coefficients used to balance different components' influence in the loss functions. Through the defined loss function, the performance of the underwater image enhancement model can be effectively evaluated and optimized, improving the quality of enhanced images. When needed, the optimal weight of the model can be loaded from the storage files for inference and further training.

3.2. Ablation Study

For the evaluation of underwater images, evaluation metrics include the Peak Signal-to-Noise Ratio (PSNR)[40], Structural Similarity Index (SSIM)[41] and the Mean Squared Error (MSE). MSE represents the mean squared error between two approximate images *I* and *K*, as defined in Equation (12):

$$MSE=\frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[I\left(i,j\right)-K\left(i,j\right)\right]^{2} \tag{12}$$

The PSNR metric represents the ratio of the maximum signal to the mean squared error of the signal. It is represented by the logarithmic decibel units, as indicated in equation (13):

$$PSNR=10*\log_{10}\left(\frac{MAX_I^2}{MSE}\right)=20*\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{13}$$

Where, MAX_I denotes the maximum value of image color. Higher PSNR values indicate clearer image. SSIM requires two input images to assess their similarity. One of images is an uncompressed and undistorted image, and the other is the restored image. So, SSIM can serve as a metric for quality assessment. Assuming x and y are the two input images, the SSIM(x, y) is defined in equation (14):

$$SSIM\left(x,y\right)=\left[l\left(x,y\right)\right]^{\alpha}\left[c\left(x,y\right)\right]^{\beta}\left[s\left(x,y\right)\right]^{\gamma} \tag{14}$$

Here, α> 0, β> 0 and γ>0. $l(x, y)$, $c(x, y)$ and $s(x, y)$ are defined in equation (15): and (16):

$$\begin{cases} l\left(x,y\right)=\dfrac{2u_x u_y+c_1}{u_x^2+u_y^2+c_1} \\ c\left(x,y\right)=\dfrac{2\sigma_{xy}+c_2}{\sigma_x^2+\sigma_y^2+c_2} \\ s\left(x,y\right)=\dfrac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3} \end{cases} \tag{15}$$

$$\begin{cases} u_x=\dfrac{1}{N}\sum\limits_{i=1}^{N}x_i \\ \delta_x=\left(\dfrac{1}{N-1}\sum\limits_{i=1}^{N}\left(x_i-\mu_x\right)^2\right)^{1/2} \\ Cov\left(X,Y\right)=E\left(X-E\left(X\right)\right)\left(Y-E\left(Y\right)\right) \end{cases} \tag{16}$$

Among them, *c1*, *c2*, and *c3* are constants, respectively. To prevent system errors due to a zero denominator, smaller values are used. In the actual calculation, it is common to assign α=β=γ= 1. *c3*=*c2*/2. σ_{xy} represents the covariance of *x* and *y*. SSIM is simplified in equation (17):

$$SSIM\left(x,y\right)=\frac{\left(2\mu_x\mu_y+c_1\right)\left(\sigma_{xy}+c_2\right)}{\left(u_x^2+u_y^2+c_1\right)\left(\sigma_x^2+\sigma_y^2+c_2\right)} \tag{17}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

The ablation study is a commonly used method in machine learning to evaluate the importance and contribution of various components in a model. By systematically removing certain parts of the model and observing the changes in performance, it is possible to identify which parts are relatively unimportant. Similar approach has been adopted by other scholars, as indicated in reference [42].

Rigorous ablation experiments were conducted to evaluate the proposed techniques. These experiments were conducted on the LSUI and UIEB datasets, evaluating three key factors: CNN features enhanced by Fourier transform, Transformer features based on PSNR attention and linear operations, and feature fusion with Fourier weights. Figure 8 illustrates the enhancement effects in each ablation experiment. In Figure 12, (b)/(c)/(d) all use the same input from (a). Additionally, Table 2 presents the comparison metrics of PSNR and SSIM for the ablation study.

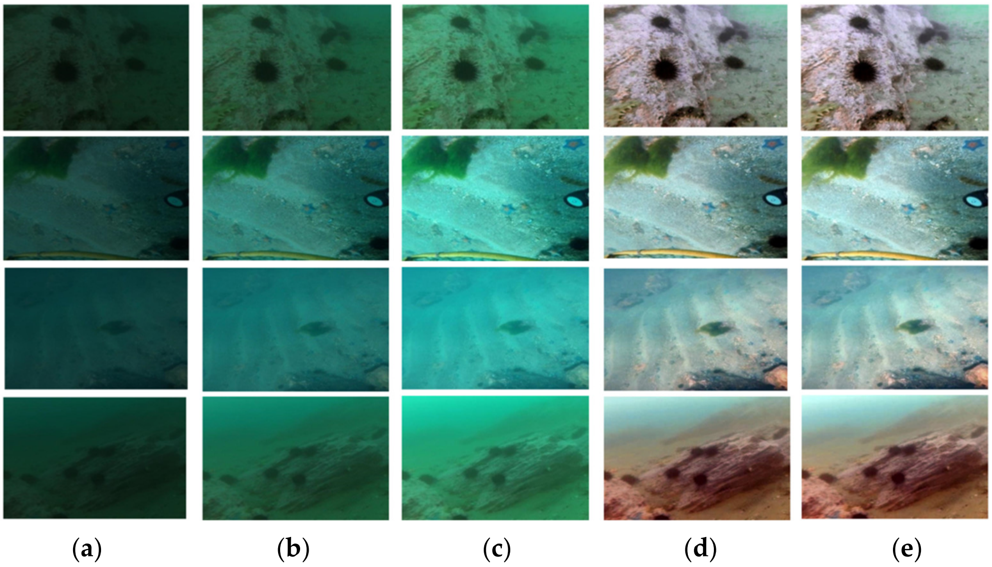


Figure 12. Ablation experiment with different components. (a) The same inputs for (b)/(c)/(d) detection methods. (b) CNN method with Fourier transform. (c) Transformer method based on PSNR attention and linear operations. (d) CNN and Transformer fusion method with Fourier weights. (e) Ground Truth.

Table 2. Comparative test of ablation experiments**.**

Structures				Fusion		LSUI		UIEB	
CNN	**Fourier**	**Transformer**	**SNR attention**	**Additive fusion**	**Fourier fusion**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
✓						15.22	0.47	13.03	0.42
✓	✓					18.82	0.64	16.77	0.60
✓	✓	✓			✓	24.83	0.79	21.70	0.70
✓	✓	✓	✓	✓		24.42	0.75	21.56	0.69
✓	✓	✓	✓		✓	26.53	0.83	23.85	0.78

Experimental results indicate that image quality can be enhanced through the utilization of CNN and Transformer architectures, respectively. Additionally, the integration of CNN and Transformer features yields a notable improvement on the image enhancement.

By utilizing the appropriate PyTorch libraries, the best trained model was loaded for test. Through normalization and resizing operations, the input images are standardized to align with preprocessing steps. Time recording tools are used to record the start and end times during the model inference. And the inference time for a single image is obtained by calculating the difference between the end and start times. In the experiment (Table 3), two types of backbone feature extraction networks are emplyed. And the times are recorded, respectively. The experiment demonstrates that element-wise Transformer attention can significantly reduce the time consumption. Additionally,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

while the dual-channel approach increases detection time, our method achieves the satisfied detection with the similar time-consumption of transformer approach.

Table 3. Comparison of different matrix-multiplication attention and computation latency**.**

Backbones	Attention	Latency(ms)
Transformer	Dot-product Transformer	2.5ms
	Element-wise Transformer	2.1ms
Transformer + CNN	Dot-product Transformer	3.0ms
	Element-wise Transformer	2.6ms

3.3. Feature Visualization Process

To validate the robustness of our feature extraction method, the feature visualization process was conducted in Figure 13. These visualized features include two types of network features. Transformer features, extracted within the Token range, improve the local perception accuracy. The Transformer network captures global and long-range features through the self-attention mechanism. The self-attention mechanism allows the Transformer to integrate features from any position within the image. This is crucial for tasks such as image restoration and color correction. According to the visualized results, it is easy to find that the Transformer can effectively restore the overall color and structure of images, overcoming the defects of CNNs in the global feature-extraction process.

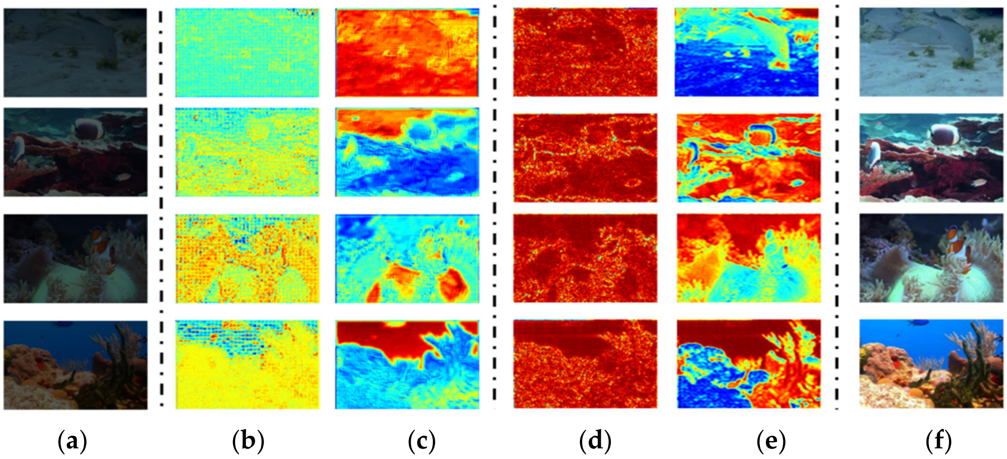


Figure 13. Visualization of the features, including the transformer branch, CNN branch and fusing weights of the original image's Fourier transform. (a) Input. (b) Transformer characters. (c) CNN characters. (d) Low pass filtering attentions. (e) High pass filtering attentions. (f) Ground truth.

Conversely, CNN features provide a global perspective, contributing to improve the global perception accuracy. By visualization, it is easy find that CNNs excel at capturing the obvious features of images. Through convolution operations, CNNs can efficiently extract image details such as edges and textures, thereby effectively suppressing noise and enhancing detail. Deeper convolutional layers enables CNNs to progressively extract high-quality features from images, which is significantly effective for removing random noise in underwater images.

Furthermore, we obtained high-pass filter and low-pass filter features by the Fourier transform, which are subsequently employed as fusion weights for the two backbones' features. High-pass filters extract edge details of image, whereas low-pass filter captures smooth information. These complementary information are multiplied with the Transformer and CNN features, respectively. This matching process enhances the accuracy of feature extraction and fusion.

Visual results show that the integration of CNNs with Transformers yields superior image enhancement effects. In summary, CNNs can remove most noise and enhance the overall color and structure, while Transformers can restore local details of the image. This combination effectively reduces noise and significantly improves the overall image quality. The effect is particularly notable when processing complicated underwater images.

Our proposed methods are compared with other methods. Figures 14(b) and (c) show the visualized global features, including the improved CNN network and the traditional ResNet network. The results indicate that the proposed method appears more prominent edge features, while the traditional ResNet method extracts relatively blurred features. The experiments demonstrate the superiority of proposed method that integrates both the time-domain and frequency-domain features.

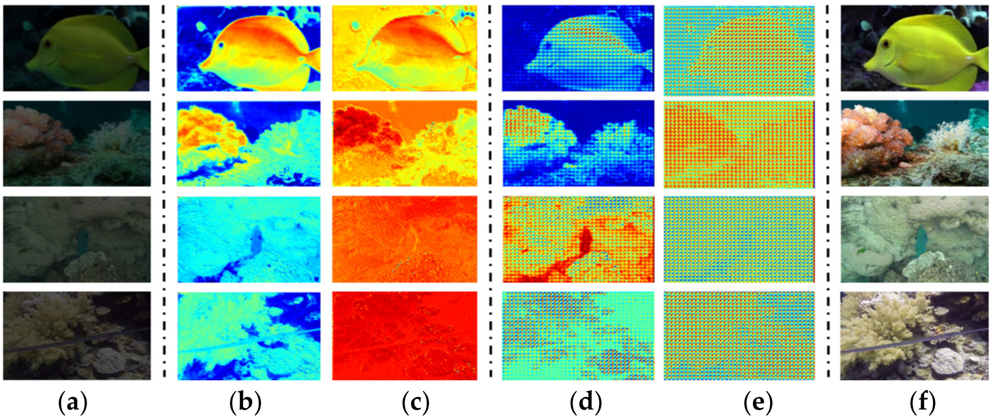


Figure 14. Validation for the proposed methods with other commonly used structures. (a) Input image. (b) The proposed CNN network incorporates both the temporal and frequency domain characters. (c) The original ResNet only adopts the temporal characters. (d) The fused feature by the Fourier transform weight of the original image. (e) The fused feature by the torch.add method. (f)Ground truth image.

Different feature-fusion methods are also compared in Figures 14(d) and (e). The results show that our method can optimize fusion weights for different objects, which enhances the feature diversity. In contrast, the torch.add method reduces the diversity and prominence of features.

3.4. Comparison with Current Methods

Our approach was qualitatively compared with other state-of-the-art (SOTA) image enhancement methods, including MIR-Net [40], U-Net [41], WaterNet [43], and Ucolor [44]. Additionally, the proposed backbone is compared quantitatively with the traditional CNN and Transformer architectures.

3.4.1. Qualitative Analysis

Visual samples of LSUI are displayed in Figure 15, which are also compared with other commonly used methods. The proposed approach demonstrates the outstanding clearance, showcasing finer details, consistent colors, and higher visibility. Additionally, the method's outputs display fewer visual artifacts, especially in zones with complicated textures.

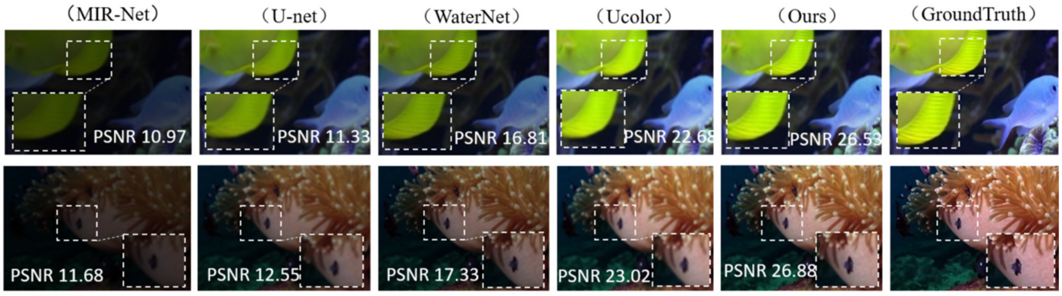


Figure 15. Qualitative analysis with the LSUI dataset.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

A visual comparison of the UIEB dataset is provided in Figure 16, highlighting our method in dealing with noisy and low-light images. The results indicate that our approach performs well in increasing image brightness, enriching image details, and suppressing noise.

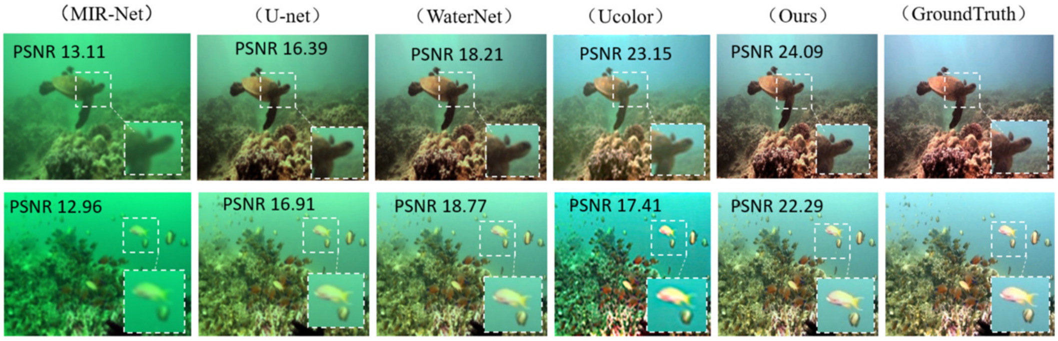


Figure 16. Qualitative analysis with the UIEB dataset**.**

3.4.2. Quantitative Analysis

In comparison to other image restoration networks, PSNR and SSIM are used to evaluate performance. Generally, higher SSIM imply the presence of more details and structure in the results. We obtained these datas from the corresponding publications or running code. All detection experiments are based on the same original input dataset, not on the optimized images from intermediate processes. Table 4 provides a comparative analysis of various methods, indicating that our algorithm outperforms others in achieving the highest PSNR and SSIM scores.

Table 4. Comparative evaluation with different image enhancement networks**.**

Methods	**LSUI**		**UIEB**	
	PSNR	**SSIM**	**PSNR**	**SSIM**
CNN[16]	15.28	0.50	13.68	0.48
MIR-Net[40]	18.80	0.66	16.78	0.63
U-net[41]	19.45	0.78	17.46	0.76
WaterNet[43]	19.62	0.80	19.27	0.83
Ucolor[44]	21.62	0.84	20.67	0.81
Transformer[26]	22.83	0.79	21.70	0.70
Ours	24.49	0.85	22.79	0.81

Compared to the Transformer method [26], it is worth noting that our linear multiplication backbone utilizes only 60% of the parameters. Additionally, in comparison with the Ucolor-based approach [44], our method demonstrates overall superiority. Furthermore, our method outperforms MIR Net [40], U-net [41], and WaterNet [43], yielding improvements of 1-3 improvement in PSNR and 0.1-0.3 in SSIM.

3.5. Comparison on Detection Tasks

To evaluate the effect of underwater image enhancement on detection tasks, the enhanced images were integrated into a series of detection algorithms, including single-stage methods SSD, RetinaNet, and GIoU [45,46]. These enhanced images were utilized as inputs for various detection tasks. The obtained detection results show that the proposed method exceeds other competing methods in detection accuracy. The visualized detection results in Figure 17 correspond with the objective outcomes, demonstrating our approach's superiority..

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

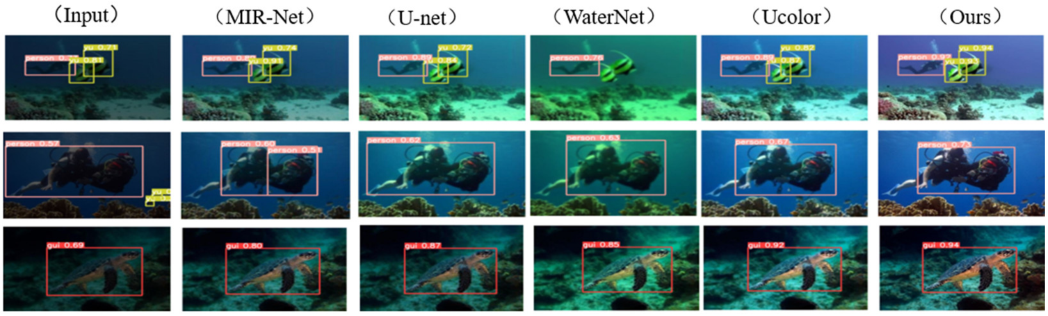


Figure 17. Visualized detection results with different image enhancement effects.

By utilizing precision-recall and recall-confidence curves as evaluation metrics, Figure 18 presents a quantitative comparison of visual detection. Due to the improved color and brightness, our method demonstrates a notable enhancement in precision and recall indexes [47]. The images enhanced by this method show superior detection outcomes, marking a significant enhancement over competing techniques.

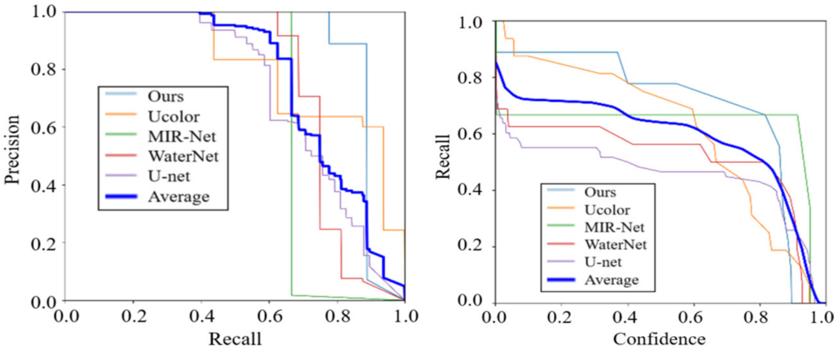


Figure 18. Precision-recall and recall-confidence curves with different image enhancement effects.

4. Conclusion

The influence of light absorption and scattering by the surrounding water leads to the loss of certain details and color information in underwater images. To address issues, such as low illumination, reduced contrast, and color shift in underwater imagery, an underwater image enhancement algorithm is proposed based on the parallel fusion of Transformer and CNN. Experiments indicate that this approach can effectively combine the global context capture ability of Transformers with the local feature extraction capability of CNNs, thereby improving the richness and accuracy of feature extraction. To effectively reduce computational load and alleviate color artifacts, a novel Transformer model integrates the PSNR attention and linear operations. Through mathematical method, this method can reduce computational complexity from $2d^2n$ to $3dn$ while simultaneously extracting constrained features. Additionally, by leveraging both temporal and frequency domain characters, a novel global feature extraction network is devised to enrich image features. The high-frequency and low-frequency information from the input image's Fourier transform are extracted, which are used to fuse different backbone's features. Experiments show that this method optimizes the fusion weights for the Transformer and CNN features, enriching the diversity of representation features. Compared to current mainstream algorithms, this method achieves optimal values in objective evaluation metrics and also produces superior subjective perceptual quality in the generated images.

Author Contributions: Conceptualization, X.L. and F.M.; methodology, X.L. and Z.C.; software, X.L.; validation, Z.C. F.M. and Z.X; formal analysis, Z.X.; investigation, Z.X. and Z.Z.; data curation, Y.W.; writing—original draft preparation, X.L.; writing—review and editing, F.M. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

Funding: This work was funded in part by Development of Intelligent Breeding Equipment for Cabin Breeding Platform (2022YFD2401104), in part by Central Public-interest Scientific Institution Basal Research Fund, FMIRI of CAFS (NO.2024YJS011), in part by the Guangdong Basic and Applied Basic Research Foundation (Grant no. 2022A1515110038), in part by the China Postdoctoral Science Foundation (NO. 2020T130474), and UMMTP-MYSP-2021 (Macau Young Scholars Program, No.AM2021003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this paper are available on request from corresponding author.

Conflicts of Interest: The author declare no conflict of interest.

References

1. Zhang, W; Liu, W; Li, L. Underwater Single-Image Restoration with Transmission Estimation Using Color Constancy. *Journal of Marine Science and Engineering*. 2022, 10(3): 430-445. doi: org/10.3390/jmse10030430.
2. Chiang, J; Chen, Y. Underwater Image Enhancement by Wavelength Compensation and Dehazing [J]. *IEEE Trans on Image Process*. 2012, 21(4): 1756-1769. doi: 10.1109/TIP.2011.2179666.
3. Hua Yang, Fei Tian, Qi Qi, Q. M. Jonathan Wu, Kunqian Li. Underwater image enhancement with latent consistency learning-based color transfer. *IET Image Processing*. 2022, 16(6): 1594-1512. doi: org/10.1049/ipr2.12433.
4. Ding, C; Dong, Lili; Xu, Wenhai. Review of histogram equalization technique for image enhancement [J]. *Computer engineering and applications*. 2017, 53(23): 12–17. doi: 10.1088/1742-6596/1019/1/012026.
5. Jingchun, Zhou; Xiaojing, Wei; Jinyu, Shi; Weishen, Chu; Weishi, Zhang. Underwater image enhancement method with light scattering characteristics. *Computers and Electrical Engineering*. 2022, 100(1): 898-915. doi: org/10.1016/j.compeleceng.2022.107898.
6. Y, Peng; X., Zhao; and P. Cosman. Single underwater image enhancement using depth estimation based on blurriness. *IEEE International Conference on Image Processing* (ICIP). 2015, 4952-4956. doi: 10.1109/ICIP.2015.7351749.
7. Song, W; Wang, Y; Huang, D. A rapid scene depth estimation model based on underwater light attenuation prior for under-water image restoration[C]. *Proceedings of 2018 Advances in Multimedia Information Processing*. 2018, 678-688. doi.org/10.1007/978-3-030-00776-8_62.
8. C., Cheng; H., Zhang; G., Li. Overview of Underwater Image Enhancement and Restoration Methods. *International Conference on CYBER Technology in Automation, Control, and Intelligent Systems* (CYBER). 2022, 520-525. doi: 10.1109/CYBER55403.2022.9907661.
9. Drews, P; Nascimento, E; Campos, M. Underwater depth estimation and image restoration based on single images[J]. *IEEE Computer Graphics and Applications*. 2016, 36(2): 24-35. doi: 10.1109/MCG.2016.26.
10. Li, J; Hou, G; Wang, G. Underwater image restoration using oblique gradient operator and light attenuation prior. *Multimed Tools Appl*. 2023, 82, 6625–6645. doi.org/10.1007/s11042-022-13605-5.
11. Ma, Z; Oh, C. A wavelet-based dual-stream network for underwater image enhancement[C]. *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2022: 2769-2773. DOI:10.1109/ICASSP43922.2022.9747781
12. Y., Zhang; Q., Jiang; P., Liu; S., Gao; X., Pan; C., Zhang. Underwater Image Enhancement Using Deep Transfer Learning Based on a Color Restoration Model. *IEEE Journal of Oceanic Engineering*. 2023, 48(2): 489-514. doi: 10.1109/JOE.2022.3227393.
13. Wang, K; Hu, Y.; Chen, J.; Wu, X.; Zhao, X.; Li, Y. Underwater Image Restoration Based on a Parallel Convolutional Neural Network. *Remote Sens*. 2019, 11, 1591-1612. doi.org/10.3390/rs11131591.
14. Y., Ueki; M., Ikehara. Underwater Image Enhancement with Multi-Scale Residual Attention Network. *International Conference on Visual Communications and Image Processing (VCIP), Munich, Germany*. 2021, pp. 1-5. doi: 10.1109/ VCIP53242. 2021. 9675342.
15. Z., Xing; M., Cai; J., Li. Improved Shallow-UWnet for Underwater Image Enhancement. *International Conference on Unmanned Systems (ICUS)*. 2022: 1191-1196. doi: 10.1109/ICUS55513.2022.9986534.
16. Chongyi, Li; Saeed, Anwar; Fatih, Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*. 2020, 98-110. doi.org/10.1016/j.patcog.2019.107038.
17. Y., Ueki; M., Ikehara. Underwater Image Enhancement with Multi-Scale Residual Attention Network. *International Conference on Visual Communications and Image Processing (VCIP), Munich, Germany*. 2021, pp. 1-5. doi: 10.1109/ VCIP53242. 2021. 9675342.
18. H., Wang; M., Yang; G., Yin; J., Dong. Self-Adversarial Generative Adversarial Network for Underwater Image Enhancement. *IEEE Journal of Oceanic Engineering*. 2024, 49(1): 237-248. doi: 10.1109/JOE.2023.3297731.

19. Y., Wang; M., J; J., Chen; J., Wu. A Novel Generative Adversarial Network for Underwater Image Enhancement. *International Conference on Intelligent Autonomous Systems (ICoIAS), Dalian, China*. 2022, pp. 84-89. doi: 10.1109/ ICoIAS56028. 2022.9931248.
20. C. Fabbri, M. J. Islam, and J. Sattar. Enhancing underwater imagery using generative adversarial networks. Proc. *IEEE Int. Conf. Robot. Autom., Brisbane, Australia.* 2018, pp. 7159–7165. doi: 10.1109/ICRA.2018.8460552.
21. G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun*. 2018, pp. 8340–8348. doi: 10.1109/CVPR.2018.00870.
22. X. Hu, M. A. Naiel, A. Wong, M. Lamm, and P. Fieguth. RUNet: A robust UNet architecture for image super-resolution. Proc. *IEEE Conf. Comput. Vis. Pattern Recognit., Long Beach, CA, USA, Jun.* 2019, pp. 505–507. doi: 10.1109/CVPRW.2019.00073.
23. Junjun Wu, Xilin Liu, Qinghua Lu. FW-GAN: Underwater image enhancement using generative adversarial network with multi-scale fusion. *Signal Processing: Image Communication*. 2022, 109:1-12. doi.org/10.1016/j.image.2022.116855.
24. Kei Terayama, Kento Shin. Integration of sonar and optical camera images using deep neural network for fish monitoring[J]. *Aquacultural Engineering*. 2019, 86: 1-7. doi.org/10.1016/j.aquaeng.2019.102000.
25. Zhang, Tianyi and Matthew Johnson-Roberson. Beyond NeRF Underwater: Learning Neural Reflectance Fields for True Color Correction of Marine Imagery[J]. *IEEE Robotics and Automation Letters*, 2023, 8(2): 6467-6474. doi.org/10.1109/ LRA. 2023. 3307287
26. Liu, Z; Lin, Y; Cao, Y. Swin Transformer: hierarchical vision transformer using shifted windows. *IEEE/CVF International Con-ference on Computer Vision. Piscataway*. 2021: 9992-10002. DOI: 10.1109/ICCV48922.2021.00986.
27. Kovács, L; Csépányi-Fürjes, L; Tewabe, W. Transformer Models in Natural Language Processing. *International Conference In-terdisciplinarity in Engineering*. 2023, Lecture Notes in Networks and Systems, vol. 929-945. doi.org/10.1007/978-3-031-54674-7_14.
28. Chang, Liu; Gang, Wang; Chen, Zhang; Pietro, Patimisco; Ruyue, Cui; Chaofan, Feng. End-to-end methane gas detection algorithm based on transformer and multi-layer perceptron. *Optics Express*. 2024, 32(1): 987-1002. DOI: 10.1364/OE.511813
29. Zamir, S; Arora, A; Khan, S. Restormer: efficient transformer for high-resolution image restoration. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: *IEEE Press*. 2022: 5718-5729. doi.org/10.48550/arXiv.2111.09881.
30. Y., Song; Z., He; H., Qian; X. Du. Vision Transformers for Single Image Dehazing. *IEEE Transactions on Image Processing*. 2023, 32, pp: 1927-1941. doi.org/10.1109/TIP.2023.3256763.
31. D, Berman; D., Levy; S., Avidan; T., Treibitz. Underwater single image color restoration using haze-lines and a new quantita-tive dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021, 43(8): 2822-2837. doi: 10.1109/ TPAMI. 2020.2977624.
32. Charu C. Aggarwal. Neural Networks and Deep Learning. *Springer*. 2018.
33. Z. Lu, X. Jiang and A. Kot. Deep Coupled ResNet for Low-Resolution Face Recognition. *IEEE Signal Processing Letters*. 2018, vol. 25, no. 4, pp. 526-530. doi: 10.1109/LSP.2018.2810121. doi: 10.1109/LSP.2018.2810121.
34. Jie, Huang; Yajing, Liu; Feng, Zhao; Keyu, Yan. Deep Fourier-Based Exposure Correction Network with Spatial-Frequency Interaction. Eur. Conf. Comput. Vis. Springer. 2022, 163–180. doi.org/10.1007/978-3-031-19800-7_10.
35. Zhou, J., Ni, J., Rao, Y. (2017). Block-Based Convolutional Neural Network for Image Forgery Detection. *Lecture Notes in Computer Science*. 2017, vol 10431. https://doi.org/10.1007/978-3-319-64185-0_6..
36. Zou, BJ., Guo, YD., He, Q. et al. 3D Filtering by Block Matching and Convolutional Neural Network for Image Denoising. J. *Comput. Sci. Technol*. 2018, 33, 838–848 (2018). https://doi.org/10.1007/ s11390-018-1859-7.
37. Abbas Shahri, A., Maghsoudi Moud, F. Landslide susceptibility mapping using hybridized block modular intelligence model. *Bull Eng Geol Environ*. 2021, 80, 267–284. https://doi.org/10.1007 /s10064-020-01922-8.
38. Q. Liu, Y. Su and P. Xu. Implementation of Artificial Intelligence Anime Styl-ization System Based on PyTorch. *Annual International Conference on Net-work and Information Systems for Computers* (*ICNISC*). 2023, pp. 84-87, doi: 10.1109/ICNISC60562.2023.00131.
39. L., Peng; C., Zhu; L., Bian. U-Shape Transformer for Underwater Image Enhancement. *IEEE Transactions on Image Processing*, vol. 32, pp. 3066-3079. 2023, doi: 10.1109/ TIP. 2023. 3276332.
40. C., Li; et al. An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Transactions on Image Processing*, vol. 29, pp. 4376-4389, 2020, doi: 10.1109/TIP.2019.2955241.
41. Basha, C; Pravallika, B; Shankar, E. An Efficient Face Mask Detector with PyTorch and Deep Learning[J]. *EAI Endorsed Transactions on Pervasive Health and Technology*. 2021, 7(25):167843. DOI:10.4108/eai.8-1-2021.167843

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2024 doi:10.20944/preprints202407.1575.v1

42. W., Li; S., Li; R., Liu. Channel Shuffle Reconstruction Network for Image Compressive Sensing. *IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates*. 2020, pp. 2880-2884. doi: 10.1109/ICIP40778.2020.9191171.
43. Zhang, Y; Liu Y, Li. Salt and pepper noise removal in surveillance video based on low-rank matrix recovery [J]. *Computational Visual Media*. 2015, 1(1): 59-68. doi.org/10.1007/s41095-015-0005-5
44. Yao, J; Liu, G. Improved SSIM image quality assessment of contrast distortion based on the contrast sensitivity characteristics of human visual system [J]. *IET Image Processing*. 2018, 12(6), 872-879. DOI:10.1049/iet-ipr.2017.0209.
45. R. Liu, Z. Jiang, S. Yang and X. Fan. Twin Adversarial Contrastive Learning for Underwater Image Enhancement and Beyond. *IEEE Transactions on Image Processing*. 2022, vol. 31, pp. 4922-4936. doi: 10.1109/TIP.2022.3190209.
46. Syed, Waqas; Aditya, Arora; Salman, Khan; Munawar, Hayat. Learning enriched features for real image restoration and en-hancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020, 45(2): 1934-1948. doi.org/10.1007/978-3-030-58595-2_30.
47. O., Ronneberger; P., Fischer; T., Brox. U-net: Convolutional network for biomedical image segmentation. Med. *Image Comput. Comput. Ass. Inter.* (MICCAI). 2015, pp. 234–241. doi.org/10.1007/978-3-319-24574-4_28.
48. Tan, L; Huang, T; Wu, L. Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification. *BMC Medical Infor-matics and Decision Making*. 2021, 324-337. doi.org/10.1186/s12911-021-01691-8.
49. Xiangyong, Liu; Guang, Chen; Xuesong, Sun; Alois, Knoll. Ground Moving Vehicle Detection and Movement Tracking Based On the Neuromorphic Vision Sensor. *IEEE Internet of Things Journal*. 2020, 7(9): 9026-9039. doi: 10.1109/ JIOT. 2020. 3001167.
50. XY., Liu; Z., Yang; J., Hou; W., Huang. Dynamic Scene's Laser Localization by NeuroIV-based Moving Objects Detection and LIDAR Points Evaluation. *IEEE Transactions on Geoscience and Remote Sensing*. 2022, doi: 10.1109 /TGRS. 2022. 3184962.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.