Li ZN, Chen XH, Guo SN *et al.* WavEnhancer: Unifying wavelet and transformer for image enhancement. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 39(2): 336-345 Mar. 2024. DOI: 10.1007/s11390-024-3414-z

WavEnhancer: Unifying Wavelet and Transformer for Image Enhancement

Zi-Nuo Li[1, †] (李梓诺), Xu-Hang Chen[1, 2, †] (陈绪行), Shu-Na Guo[1] (郭淑娜)
Shu-Qiang Wang[2, *] (王书强), *Senior Member, CCF, IEEE*
and Chi-Man Pun[1, *] (潘治文), *Senior Member, IEEE*

[1] *Department of Computer and Information Science, University of Macau, Macao 999078, China*

[2] *Research Center for Biomedical Information Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

E-mail: zinuo.li@research.uwa.edu.au; yc17491@umac.mo; sguo0039@student.monash.edu; sq.wang@siat.ac.cn
cmpun@umac.mo

Received May 19, 2023; accepted January 6, 2024.

Abstract Image enhancement is a widely used technique in digital image processing that aims to improve image aesthetics and visual quality. However, traditional methods of enhancement based on pixel-level or global-level modifications have limited effectiveness. Recently, as learning-based techniques gain popularity, various studies are now focusing on utilizing networks for image enhancement. However, these techniques often fail to optimize image frequency domains. This study addresses this gap by introducing a transformer-based model for improving images in the wavelet domain. The proposed model refines various frequency bands of an image and prioritizes local details and high-level features. Consequently, the proposed technique produces superior enhancement results. The proposed model's performance was assessed through comprehensive benchmark evaluations, and the results suggest it outperforms the state-of-the-art techniques.

Keywords transformer, wavelet transform, image enhancement

1 Introduction

In the past decade, the field of digital photography has seen remarkable growth and development, largely due to the significant strides made in camera sensor technology. This technological evolution has not only improved the quality of photographs, but also has expanded the potential for creativity and innovation in this medium. However, despite these advancements, there remain substantial challenges in the area of image enhancement, specifically in the realm of post-processing techniques. Professional software applications such as Adobe Photoshop provide a range of interactive and semi-automated capabilities that enable users to make a plethora of modifications to their photographs. However, these tools often necessitate a high level of skill and technical expertise to be used effectively, which can be a barrier for many users. The complexity inherent in these software applications may make manual adjustments a daunting task, especially for amateur photographers who may lack the necessary technical prowess or the aesthetic acuity to retouch their photographs successfully.

The challenges associated with manual image en-

Regular Paper

Special Section of CGI 2023

This work was supported in part by the Science and Technology Development Fund of Macau SAR of China under Grant Nos. 0087/2020/A2 and 0141/2023/RIA2, the National Natural Science Foundation of China under Grant No. 62172403, and the Distinguished Young Scholars Fund of Guangdong Province of China under Grant No. 2021B1515020019.

[†] Co-First Author (Zi-Nuo Li was responsible for the conceptualization, data curation, and writing of the original draft. Xu-Hang Chen contributed to the methodology, visualization, and editing of the manuscript.)

[*] Corresponding Author (Shu-Qiang Wang conducted the formal analysis and contributed to reviewing and editing the manuscript. Chi-Man Pun was involved in reviewing and editing the paper, as well as acquiring the necessary funding for the research.)

hancement have led to the development of fully automated strategies. The said strategies aim to replace non-expert users' work or offer experienced artists an improved starting point for manual editing. In image enhancement, photographers frequently use both local filters and global modifications in combination. However, learning-based automatic image enhancement is now being optimized with neural networks due to the emergence of deep learning. These learning-based models offer the promise of greatly improved results. By training neural networks on large datasets of professionally retouched photographs, these algorithms can learn the intricate, nuanced adjustments that expert photographers make. They can then apply these learned techniques to new photographs, effectively replicating the skills of a professional. This opens up new possibilities for non-expert users, who can now achieve high-quality results without the need of extensive technical skills or aesthetic judgment. A large number of excellent studies have emerged, such as [1–9].

Although image enhancement has been extensively studied, there are still significant challenges. First, certain studies limit themselves to modifying individual pixels or the image as a whole, which can overlook both the global tone and subtle changes as shown in Fig.1. Second, the optimization of images using diverse frequency priors is rare, resulting in suboptimal outcomes.

To tackle these challenges, this study derives inspiration from two sources: the wavelet transform and the vision transformer (ViT)[12]. The method aims to extract new features from photos and enhance them in various frequency domains, which allows the network to extract information from different frequency subbands and improve the model's ability to handle various image structures and patterns. In this case, we can improve the overall quality of the enhanced image by capturing more detailed and accurate information from the input image.

This study's primary contributions are as follows.

1) Inadequate research conducted solely on the pixel or global level yields unsatisfactory outcomes. This work proposes the WavEnhancer, a novel framework based on the wavelet domain transformer. The WavEnhancer framework places emphasis on both pixel and global levels.

2) In this study, we propose a model that combines multi-frequency and global refinement techniques. We evaluate its superior performance compared with the state-of-the-art methods using public benchmark datasets MIT-Adobe FiveK[13] and HDR+[14].

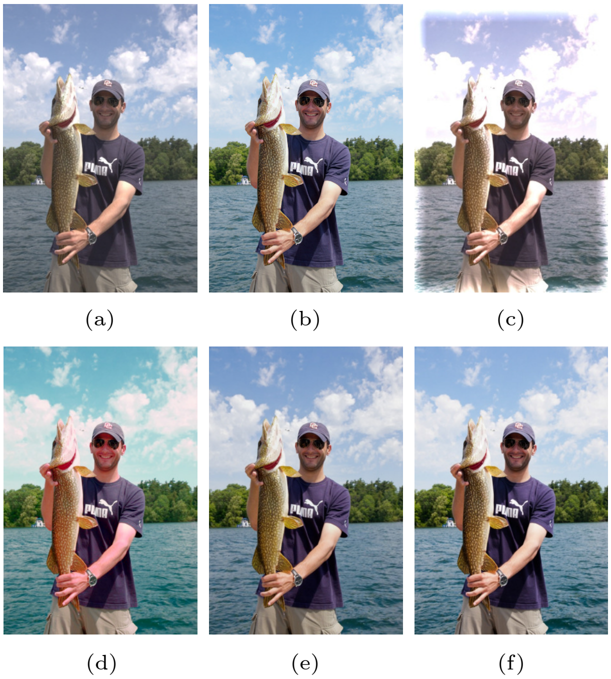


Fig.1. This photograph displays both (a) an input image and (f) a target image. For white balance and exposure, (b) DPE[10], (c) UPE[11], and (d) CSRNet[5] produce results that are inconsistent with the ground truth. (e) The image we enhanced is closer to the target.

The remainder of this paper is organized as follows. In Section 2, we present a comprehensive review of related work, focusing on image enhancement and the application of wavelet transforms in this domain. Section 3 describes our proposed method in detail, including the mathematical formulation and the algorithmic steps involved. In Section 4, we present our experimental setup, discuss the results obtained, and provide a comparative analysis with state-of-the-art methods. We also discuss the limitations of our model and potential future directions. Finally, in Section 5, we conclude the paper by summarizing our contributions and highlighting the significance of our work in the context of image enhancement and wavelet-based methods.

2 Related Work

2.1 Image Enhancement

In recent years, deep learning has emerged as a powerful contender in image enhancement. Many image enhancement studies have focused on either local or global aspects.

Several methods aim at improving local image refinement. For instance, UPE (Underexposed Photo Enhancement)[11] uses an encoder-decoder architecture to detect scaling luminance maps, while DPE (Deep Photo Enhancer)[10] creates intermediate lighting connections to predict enhancement outcomes. HDRNet[15] applies bilateral grid processing and local affine color transforms, whereas CSRNet (Conditional Sequential Modulation Network)[5] is a lightweight retouching framework. DeepLPF (Deep Local Parametric Filters)[1] trains spatially local filters to enhance images.

Other methods focus on enhancing the overall image. Bychkovsky *et al.*[13] created the MIT-Adobe FiveK dataset and employed regression-based techniques to detect photographers' alterations in image pairs. STAR-DCE[8] introduces a lightweight transformer network that enhances real-time image quality. 3D-LUT[7] learns 3D look-up tables (LUTs) using annotated data, whereas SepLUT[6] decomposes a color transformation into component-independent and component-correlated sub-transformations.

2.2 Wavelet Transform

The wavelet transform is a time-tested, conventional technique that enables images to be downsampled and upsampled without loss. The transform also enables multi-frequency refinement.

Deep learning has rapidly advanced, leading to a combination of wavelet transform with deep learning in numerous studies. As an example, WCT^2 (Wavelet Corrected Transfer Based on Whitening and Coloring Transforms)[16] preserves structural information and statistical properties when stylizing features in the latent feature space. MWCNN (Multi-Level Wavelet Convolutional Neural Network)[17] integrates wavelet transform into the convolutional neural network (CNN) architecture to reduce the feature map resolution and increase the receptive field simultaneously. FP-GAN (Fine Perceptive Generative Adversarial Network)[18] produces high-resolution images in multi-frequency by converting low-resolution magnetic resonance images. Wave-ViT[19] unites invertible downsampling with wavelet transforms and self-attention learning to achieve self-attention learning with lossless downsampling.

This work is inspired by the previous studies mentioned and proposes a novel methodology to optimize images using a combination of wavelet and transformer techniques for multi-frequency refinement and low-cost downsampling. It is important to note that the typical Haar wavelet transform is the technique used in our model due to its ability to split the initial image into distinct channels that encapsulate various elements, thereby facilitating enhanced stylization[16, 20]. By utilizing multi-frequency refinement, we enhance images at both the local and global levels, thereby achieving superior results.

3 Methodology

3.1 Overall Framework

The proposed image enhancement model comprises of three main components, namely the wavelet transform, a global stylization remapping module (GSR), and a detailed parametric refinement module (DPR). The entire workflow of the proposed model is illustrated and presented in Fig.2. Specifically, our method uses multi-frequency feature extraction to obtain richer information and then uses global refinement to fuse this information together to further improve the restoration quality.

Wavelet transform is an extensively used technique for downsampling images efficiently at minimal computational cost while ensuring no loss of information. This makes it highly suitable for our multi-frequency optimization method.

The initial stage in our model involves subjecting the input image to a discrete wavelet transformation (DWT) process that results in a low-low (LL) subband channel representing the low-frequency region as well as high-low (HL), low-high (LH), and high-high (HH) subband channels describing high-frequency sectors. This process retains approximation coefficients, which represent the geometric features and color context of the low-frequency regions, within the LL channel. Meanwhile, the high-frequency regions retrieve textural information from the HL, LH, and HH channels.

The high-frequency channels undergo processing using U-Net blocks[21, 22] with Smooth L1 regularization to enable convergence, while the low-frequency region serves as an input for the GSR module. Subsequently, both modules generate their corresponding refined components, and these outputs are integrated using the inverse discrete wavelet transformation (IDWT) technique to reconstruct the image. Finally, we introduce our detailed parametric refinement module to produce an enhanced stylized output.

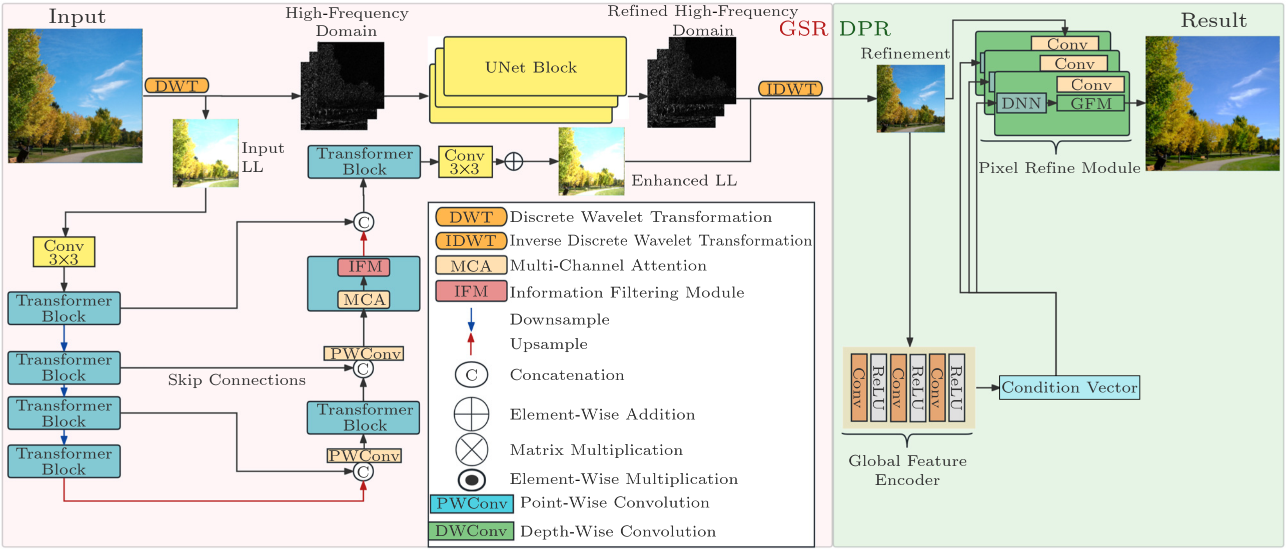


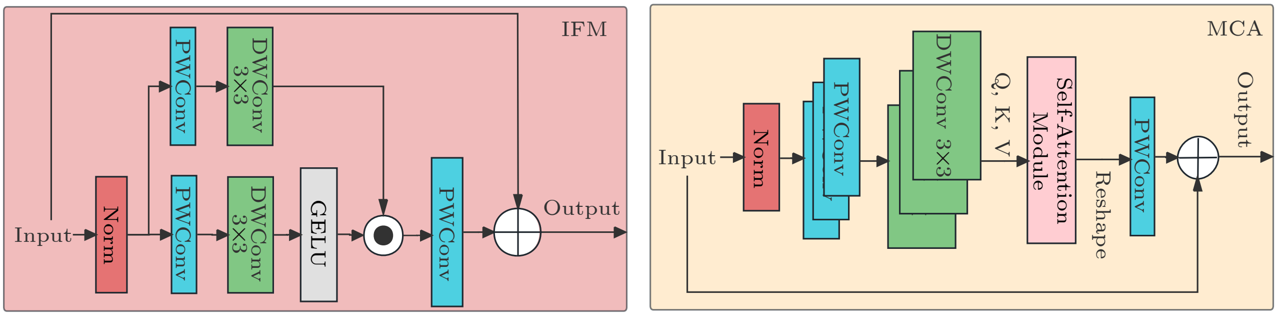
Fig.2. Our WavEnhancer's architecture has three components: the wavelet transform, a global stylization remapping module (GSR) for global feature restoration, and a detailed parametric refinement module (DPR) for visual quality improvement.

3.2 Global Stylization Remapping

The use of global feature encoder in GSR simplifies the integration of global color information, resulting in improved picture enhancement performance. The LL band derived from the wavelet transformer serves as a summary of the image content, retaining all of the original image's content information. To stylize and remap this component, we adopt the Restormer[23] and utilize transformer blocks. The self-attention module of the transformer is effective in gathering global information, making it highly suitable for the LL component, which requires color and texture refinement.

In the initial stage of the LL refinement process, a 3×3 convolution is applied to the LL features, resulting in an output shape of $(H/2) \times (W/2) \times C$, where H represents height, W represents width, and C represents channel. Four symmetric encoder-decoder transformer blocks are used subsequently to convert the LL features into deep features. To ensure efficiency, the number of transformer blocks increases from top to bottom. Conversely, pixel reshuffle enables the up and down sampling among the four blocks. Each transformer block incorporates the information filtering module (IFM) and multi-channel attention module (MCA) as shown in Fig.3. To reduce computational effort, MCA is employed in place of the conventional self-attention, which involves significant computational effort. At each level, the IFMs govern the flow of information, enabling minor aspects to be focused on and complementing the other levels.

The MCA calculates channel-wise attention to encode global contextual information implicitly, rather than spatially. First, a normalized layer tensor $\boldsymbol{Y} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ is subjected to depth-wise convolution operations to generate $\boldsymbol{Q}$ (query), $\boldsymbol{K}$ (key), and $\boldsymbol{V}$ (value). These computations enable self-attentive maps that emphasize local information. Pixel-wise cross-channel context is aggregated using 1×1 convo-



(a)

(b)

Fig.3. (a) IFM and (b) MCA modules.

lutions, and then 3×3 depth-wise convolutions encode channel-wise spatial context, accomplishing MCA. Here, $\boldsymbol{Q}=W_{\mathrm{d}}^{Q}W_{\mathrm{p}}^{Q}\boldsymbol{Y}$, $\boldsymbol{K}=W_{\mathrm{d}}^{K}W_{\mathrm{p}}^{K}\boldsymbol{Y}$, and $\boldsymbol{V}=W_{\mathrm{d}}^{V}W_{\mathrm{p}}^{V}$, with $W_{\mathrm{p}}^{(\cdot)}$ and $W_{\mathrm{d}}^{(\cdot)}$ denoting the 1×1 point-wise convolution and 3×3 depth-wise convolution, respectively. Subsequently, the query and key projections are reshaped, and their dot-product interaction produces a transposed-attention map $\boldsymbol{A}$ of size $\mathbb{R}^{\hat{C}\times\hat{C}}$, resulting in the formulation of MCA as (1) and (2).

$$\hat{\boldsymbol{X}}=W_{\mathrm{p}}\times Attention(\hat{\boldsymbol{Q}},\hat{\boldsymbol{K}},\hat{\boldsymbol{V}})+\boldsymbol{X}, \tag{1}$$

$$Attention(\hat{\boldsymbol{Q}},\hat{\boldsymbol{K}},\hat{\boldsymbol{V}})=\hat{\boldsymbol{V}}\cdot softmax(\hat{\boldsymbol{K}}\cdot\hat{\boldsymbol{Q}}/\alpha), \tag{2}$$

where the input and output feature maps are represented by $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, respectively. We reshape tensors from their original dimensions of $\mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ to obtain $\hat{\boldsymbol{Q}}\in\mathbb{R}^{\hat{H}\hat{W}\times\hat{C}}$, $\hat{\boldsymbol{K}}\in\mathbb{R}^{\hat{C}\times\hat{H}\hat{W}}$, and $\hat{\boldsymbol{V}}\in\mathbb{R}^{\hat{H}\hat{W}\times\hat{C}}$. A learnable scaling parameter α controls the magnitude of the dot product of $\boldsymbol{K}$ and $\boldsymbol{Q}$ before applying the $softmax$ function. Similar to the conventional multi-head self-attention methods[12, 24], we divide the number of channels into heads and simultaneously learn different attention maps.

The *Gating* mechanism is elaborated in (3).

$$Gating(\boldsymbol{X})=\phi\left(W_{\mathrm{d}}^{1}W_{\mathrm{p}}^{1}(LN(\boldsymbol{X}))\right)\odot W_{\mathrm{d}}^{2}W_{\mathrm{p}}^{2}(LN(\boldsymbol{X})), \tag{3}$$

where $\odot$ denotes the element-wise multiplication, ϕ represents the $GELU$ non-linearity, and LN stands for layer normalization[25]. IFM has a unique role compared with MCA, which focuses on contextually enhancing features. Considering that the proposed IFM implements more operations than the regular feed-forward network[12], we reduce the expansion ratio to ensure that the number of parameters and computational expense are consistent.

To complete the process, the LL component is restored via a 3×3 convolution, resulting in a shape of $(H/2)\times(W/2)\times C$. We define the refinement loss as L_{R} through (4).

$$L_{\mathrm{R}}=\sum_{i=1}^{N}\left\{\omega_{\mathrm{Lab}}\left\|Lab\left(\hat{Y}_i\right)-Lab\left(Y_i\right)\right\|_1+\omega_{\text{MS-SSIM}}MS\text{-}SSIM\left(L\left(\hat{Y}_i\right),L\left(Y_i\right)\right)\right\}, \tag{4}$$

where the ground truth LL is represented by Y_i, and the enhanced LL is denoted by $\hat{Y}_i$. Here, $Lab(x)$ returns CIELab channels that correspond to the RGB channels in the original images, whereas $L(x)$ returns the image's CIELab $\boldsymbol{L}$ channel. $MS\text{-}SSIM$ stands for multi-scale structural similarity function, and the hyperparameters ω_{Lab} and $\omega_{\text{MS-SSIM}}$ indicate the relative importance of different components in the loss function.

The next step involves merging the enhanced LL and the refined high-frequency domain via IDWT, which then generates a refined intermediate result.

3.3 Detailed Parametric Refinement

Our model design is inspired by CSRNet[5]. The intermediate result from the preceding stage is fed into our detailed parametric refinement module for improved processing. Within the DPR, the pixel refine module takes in the low-quality image and generates the stylized image. In parallel, the global feature encoder estimates priors based on the input image and controls the pixel refine module by means of global feature modulation (GFM) operations.

The pixel refine module is a fully convolutional architecture that consists of N layers with $N\times1$ ReLU activations. The module's unique property is that all filter sizes are 1×1, enabling individual manipulation of each pixel in the input image. Thus, the pixel refine module processes each pixel independently and glides across the input image. The global feature encoder contains three blocks: convolution, ReLU, and downsampling layers, to capture global information. The output of the global feature encoder is a condition vector that is subsequently fed into the pixel refine module.

GFM is a variant of AdaFM[26]. When the filter g_i has dimensions of 1×1, AdaFM morphs into GFM. Additionally, as demonstrated in (5), GFM can scale and shift the feature map x_i using affine parameters γ and β without normalizing it. This leads to a more effective adaptation of the feature map to the specific characteristics of the input image.

$$GFM(x_i)=\gamma\times x_i+\beta. \tag{5}$$

3.4 Objective Function

To achieve an ideal color effect in diverse contexts via feature-level information, we need to compare the actual picture's feature with the feature yielded by the generated image for closer high-level information. The prevalent perceptual loss[27, 28] measures overall perception in low-level computer vision tasks and has proved to be effective because it uses specific layers of a pre-trained VGG-16. Thus, in our implementation, we employ perceptual loss and VGG-16 as the backbone to ensure model convergence.

$$L_{\Phi} = \sum_{k=0}^{5} \lambda_l ||\Phi_k(I^{'}) - \Phi_k(I)||_1, \tag{6}$$

where $I^{'}$ and I indicate the reference image and input image, respectively. We compare the differences between CONVk2 ($k = 1, \ldots, 5$) and the original images ($k = 0$ in (6)) with respect to the ground truth and the enhanced image. The VGG16 model, Φ, pre-trained on the ImageNet dataset is used to evaluate the differences. Furthermore, we use smooth L1 loss ($\text{smooth}_{\text{L}_1}$) and L_{R} to constrain our pixel-wise enhancement model. The refinement loss L_{R} is defined as (4).

The final loss function of our model, which incorporates all regularization items, is presented in (7):

$$L_{\text{total}} = L_{\Phi} + \lambda_{\text{R}}\lambda_{\text{R}} + \lambda_{\text{smooth}_{\text{L}_1}} \text{smooth}_{\text{L}_1}. \tag{7}$$

Here, we utilize two constant parameters, λ_{R} and $\lambda_{\text{smooth}_{\text{L}_1}}$, to govern the effects of features and pixel regularization terms, respectively. The values of these parameters are set empirically as $\lambda_{\text{R}} = 2$ and $\lambda_{\text{smooth}_{\text{L}_1}} = 2$ to achieve optimal results.

4 Experiments and Results

4.1 Experimental Setup

We use the MIT-Adobe FiveK[13] and HDR+[14] datasets for training and evaluation purposes. The MIT-Adobe FiveK dataset, which contains five retouched copies of 5 000 original photos from a variety of contexts, is currently the largest image enhancement dataset. The HDR+ dataset, intended for the high-dynamic range and low-light imagery captured using Google Camera burst photography, consists of 3 640 scenes. To ensure a fair comparison, we follow the same dataset configuration as 3D-LUT[7] and transform all images in the standard PNG format and a resolution of 480p.

4.2 Implementation Detail

We implement our model using PyTorch on an RTX A6000. To train the model, we employ the standard Adam Optimizer with default settings, with a learning rate of 1×10^{-4} and a batch size of 1, and set the number of epoch to 200. Additionally, we utilize data augmentation techniques such as random cropping, horizontal flipping, as well as brightness and saturation adjustments. All the images are resized to 512×512 at the training stage and tested on the original size at the inference stage. The number of transformer blocks is set to 6, and the number of U-Net blocks is set to 3, which is consistent with the number of high-frequency components.

4.3 Evaluation Metrics

We assess the effectiveness of various methods using three different metrics: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and ΔE. PSNR is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. The SSIM index is the measurement or prediction of image quality based on an initial uncompressed or distortion-free image as reference. The ΔE metric evaluates color variation perceived by the human eye within the CIELab color space[29]. Higher values for both PSNR and SSIM indicate superior performance, while a lower value for ΔE indicates a more visually appealing color.

4.4 Quantitative Comparisons

We benchmark our method with eight state-of-the-art methods: HDRNet[15], DPE[10], UPE[11], CSRNet[5], DeepLPF[1], 3D-LUT[7], STAR-DCE[8], and SepLUT[6]. The results of our quantitative evaluations, which includes PSNR, SSIM, and ΔE metrics, are displayed in Table 1. The notation "↑" indicates that a larger value implies better metrics, while "↓" indicates that a lower value implies better metrics. In cases where a result is not available, it is marked as "N/A". The top-performing result is highlighted in bold. Our method outperforms all the others in all metrics, as displayed in the table. While possible, we

Table 1. Quantitative Comparisons of Various Image Enhancement Methods on the MIT-Adobe FiveK and HDR+ Datasets

Method	FiveK			HDR+		
	PSNR ↑	SSIM ↑	ΔE ↓	PSNR ↑	SSIM ↑	ΔE ↓
HDRNet	19.93	0.798	14.42	23.04	0.879	8.97
DPE	17.66	0.725	17.71	22.56	0.872	10.45
UPE	21.88	0.853	10.80	21.21	0.816	13.05
CSRNet	17.85	0.790	18.27	N/A	N/A	N/A
DeepLPF	24.55	0.846	8.62	N/A	N/A	N/A
3D-LUT	24.59	0.846	8.30	23.54	0.885	7.93
STAR-DCE	24.50	0.893	N/A	26.50	0.883	5.77
SepLUT	25.02	0.873	7.91	N/A	N/A	N/A
Ours	**25.46**	**0.896**	**7.28**	**28.68**	**0.905**	**4.89**

perform an additional evaluation of each model using their available pre-trained models on both datasets.

Fig.4 and Fig.5 present the qualitative results obtained by our method in terms of landscape and por-

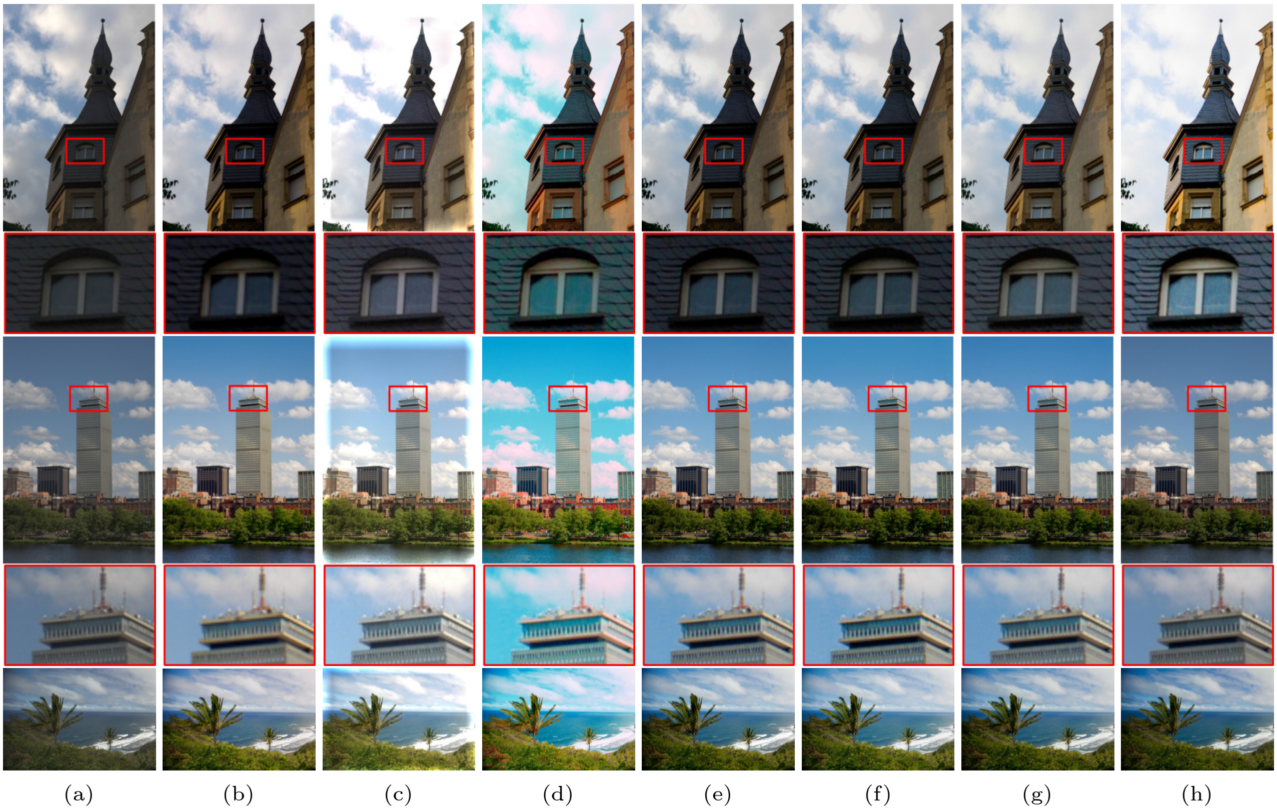


Fig.4. We conduct a visual comparison of different image enhancement methods for natural scenes and constructions. The visual results are pleasing and satisfactory. (a) Input. (b) DPE[10]. (c) UPE[11]. (d) CSRNet[5]. (e) 3D-LUT[7]. (f) SepLUT[6]. (g) Ours. (h) Target.

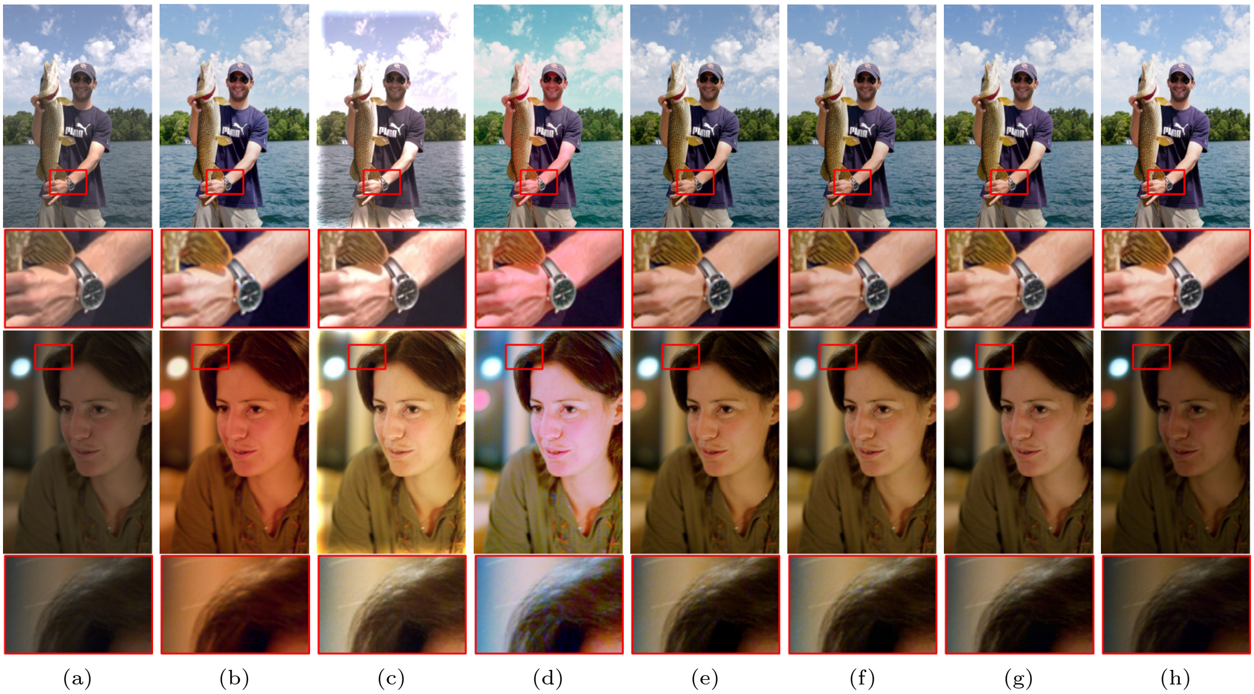


Fig.5. Our portrait retouching result yields substantial improvements. (a) Input. (b) DPE[10]. (c) UPE[11]. (d) CSRNet[5]. (e) 3D-LUT[7]. (f) SepLUT[6]. (g) Ours. (h) Target.

trait scenario, respectively. In Fig.4, it can be seen that our method outperforms the methods of DPE, UPE, and CSRNet in these areas, whose results differ significantly from the desired outcome. Specifically, the color, exposure, and detail reproduction performances of the aforementioned methods do not meet high quality. Comparatively, 3D-LUT and SepLUT produce superior outcomes, although their tone mapping is not completely satisfactory, causing the resulting images to be either too bright or too dark compared with the target. Our model underlines the importance of selecting the most effective image enhancement method to achieve superior image quality. In Fig.5, it can be observed that our method can recover finer details, like sharpness of hair and watches, which tend to vanish in prior techniques. Furthermore, our tone mapping generates visually similar results to the ground truth.

4.5 Ablation Studies

In this subsection, we analyze our ablation studies to investigate the impact and selection of our modules. Table 2 displays the acronyms for our GSR and DPR. The top-performing result is highlighted in bold. We offer two alternative modules to GSR and DPR: ConvNet[30] and U-Net[21]. ConvNet is a popular cascaded aggregation network that has been shown to be an effective substitute for GSR in image restoration, while U-Net uses the equivalent downsampling pattern of GSR.

Table 2. Ablation Study on the MIT-Adobe FiveK and HDR+ Datasets

Method	FiveK			HDR+		
	PSNR ↑	SSIM ↑	ΔE ↓	PSNR ↑	SSIM ↑	ΔE ↓
GSR+UNet	24.28	0.874	8.57	24.17	0.823	8.89
ConvNet+DPR	24.40	0.873	8.43	26.84	0.867	6.34
ConvNet+UNet	21.43	0.815	12.23	25.85	0.860	7.28
UNet+DPR	24.06	0.879	9.42	25.48	0.862	7.87
UNet+UNet	22.04	0.838	11.72	23.87	0.840	9.49
Ours	**25.46**	**0.896**	**7.28**	**28.68**	**0.905**	**4.89**

The first half of the ablation experiments aims to enhance the low-frequency component, whereas the second half focuses on global optimization. We do not alter our U-Net block for refining high-frequency domains.

We first attempt to replace DPR with U-Net for global optimization, but this leads to a decline in metrics. Subsequently, we replace GSR with ConvNet and use both DPR and U-Net, but metrics are not so strong as our model, and significantly decline with ConvNet and U-Net. Additionally, we try pairing U-Net with both DPR and U-Net after replacing GSR, but the results are not so good as our full model. Therefore, our study shows that the most effective combination for achieving the highest metrics and best visual impression is GSR and DPR.

4.6 Limitations

Our proposed neural network has limitations that must be considered. Firstly, the network includes a vast number of parameters, which prolongs training time, leading to a significant challenge. This is particularly problematic in real-time systems where processing speed is critical. Secondly, while our model demonstrates promising outcomes on various datasets, some discrepancies still exist between our results and the ground truth. In some instances, the network's outcomes are over-enhanced, resulting in excessive brightness and inadequate detail.

The highlighted challenges emphasize the importance of conducting additional research to optimize the efficiency and precision of deep learning based image enhancement methods. Moreover, identifying these challenges can provide direction for future research to experiment with alternative network frameworks and loss functions that may help overcome these limitations.

5 Conclusions

This paper proposed a new model for improving the different frequency bands of an image. The method employs a transformer-based model that operates within the wavelet domain, combining DWT modules with transformer modules to optimize the low-frequency region of the image. The IDWT produced by the transformer undergoes additional processing through U-Net's optimized high-frequency domain before being fed into our global stylization remapping module for further improvement. Our method emphasizes not only regional but also global optimization, setting it apart from other state-of-the-art methods. As a future direction, we seek to enhance performance further by increasing the downsampling multiplier of wavelet pooling and incorporating an attention mechanism into the model.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] Moran S, Marza P, McDonagh S, Parisot S, Slabaugh G. DeepLPF: Deep local parametric filters for image enhancement. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.12826–12835. DOI: 10.1109/cvpr42600.2020.01284.

[2] Li Z N, Chen X H, Wang S Q, Pun C M. A large-scale film style dataset for learning multi-frequency driven film enhancement. In *Proc. the 32nd International Joint Conference on Artificial Intelligence*, Aug. 2023, pp.1160–1168. DOI: 10.24963/ijcai.2023/129.

[3] Li Z N, Chen X H, Pun C M, Cun X D. High-resolution document shadow removal via a large-scale real-world dataset and a frequency-aware shadow erasing net. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision*, Oct. 2023, pp.12415–12424. DOI: 10.1109/iccv51070.2023.01144.

[4] Luo S H, Chen X H, Chen W W, Li Z N, Wang S Q, Pun C M. Devignet: High-resolution vignetting removal via a dual aggregated fusion transformer with adaptive channel expansion. arXiv: 2308.13739, 2023. https://arxiv.org/abs/2308.13739, Mar. 2024.

[5] He J W, Liu Y H, Qiao Y, Dong C. Conditional sequential modulation for efficient global image retouching. In *Proc. the 16th European Conference on Computer Vision*, Aug. 2020, pp.679–695. DOI: 10.1007/978-3-030-58601-0_40.

[6] Yang C Q, Jin M G, Xu Y, Zhang R, Chen Y, Liu H D. SepLUT: Separable image-adaptive lookup tables for real-time image enhancement. In *Proc. the 17th European Conference on Computer Vision*, Oct. 2022, pp.201–217. DOI: 10.1007/978-3-031-19797-0_12.

[7] Zeng H, Cai J R, Li L D, Cao Z S, Zhang L. Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2020, 44(4): 2058–2073. DOI: 10.1109/tpami.2020.3026740.

[8] Zhang Z Y, Jiang Y T, Jiang J, Wang X G, Luo P, Gu J W. STAR: A structure-aware lightweight transformer for real-time image enhancement. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision*, Oct. 2021, pp.4086–4095. DOI: 10.1109/iccv48922.2021.00407.

[9] Hu S Y, Yu W, Chen Z, Wang S Q. Medical image reconstruction using generative adversarial network for Alzheimer disease assessment with class-imbalance problem. In *Proc. the 6th International Conference on Computer and Communications*, Dec. 2020, pp.1323–1327. DOI: 10.1109/iccc51575.2020.9344912.

[10] Chen Y S, Wang Y C, Kao M H, Chuang Y Y. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.6306–6314. DOI: 10.1109/cvpr.2018.00660.

[11] Wang R X, Zhang Q, Fu C W, Shen X Y, Zheng W S, Jia J Y. Underexposed photo enhancement using deep illumination estimation. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.6842–6850. DOI: 10.1109/cvpr.2019.00701.

[12] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. the 9th International Conference on Learning Representations*, May 2021.

[13] Bychkovsky V, Paris S, Chan E, Durand F. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp.97–104. DOI: 10.1109/cvpr.2011.5995332.

[14] Hasinoff S W, Sharlet D, Geiss R, Adams A, Barron J T, Kainz F, Chen J W, Levoy M. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Trans. Graphics*, 2016, 35(6): Article No. 192. DOI: 10.1145/2980179.2980254.

[15] Gharbi M, Chen J W, Barron J T, Hasinoff S W, Durand F. Deep bilateral learning for real-time image enhancement. *ACM Trans. Graphics*, 2017, 36(4): Article No. 118. DOI: 10.1145/3072959.3073592.

[16] Yoo J, Uh Y, Chun S, Kang B, Ha J W. Photorealistic style transfer via wavelet transforms. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27–Nov. 2, 2019, pp.9035–9044. DOI: 10.1109/iccv.2019.00913.

[17] Liu P J, Zhang H Z, Lian W, Zuo W M. Multi-level wavelet convolutional neural networks. *IEEE Access*, 2019, 7: 74973–74985. DOI: 10.1109/access.2019.2921451.

[18] You S R, Lei B Y, Wang S Q, Chui C K, Cheung A C, Liu Y, Gan M, Wu G C, Shen Y Y. Fine perceptive GANs for brain MR image super-resolution in wavelet domain. *IEEE Trans. Neural Networks and Learning Systems*, 2023, 34(11): 8802–8814. DOI: 10.1109/TNNLS.2022.3153088.

[19] Yao T, Pan Y W, Li Y H, Ngo C W, Mei T. Wave-ViT: Unifying wavelet and transformers for visual representation learning. In *Proc. the 17th European Conference on Computer Vision*, Oct. 2022, pp.328–345. DOI: 10.1007/978-3-031-19806-9_19.

[20] Liu L, Liu J Z, Yuan S X, Slabaugh G, Leonardis A, Zhou W G, Tian Q. Wavelet-based dual-branch network for image demoiréing. In *Proc. the 16th European Conference on Computer Vision*, Aug. 2020, pp.86–102. DOI: 10.1007/978-3-030-58601-0_6.

[21] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Oct. 2015, pp.234–241. DOI: 10.1007/978-3-319-24574-4_28.

[22] Hu S Y, Yuan J P, Wang S Q. Cross-modality synthesis from MRI to PET using adversarial U-Net with different normalization. In *Proc. the 2019 International Conference on Medical Imaging Physics and Engineering*, Nov. 2019, pp.1–5. DOI: 10.1109/icmipe47306.2019.9098219.

[23] Zamir S W, Arora A, Khan S, Hayat M, Khan F S, Yang M H. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun.

2022, pp.5718–5729. DOI: 10.1109/cvpr52688.2022.00564.

[24] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In *Proc. the 31st Conference on Neural Information Processing Systems*, Dec. 2017, pp.6000–6010. DOI: 10.5555/3295222.3295349.

[25] Ba J L, Kiros J R, Hinton G E. Layer normalization. arXiv: 1607.06450, 2016. https://arxiv.org/abs/1607.06450, Mar. 2024.

[26] He J W, Dong C, Qiao Y. Modulating image restoration with continual levels via adaptive feature modification layers. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.11048–11056. DOI: 10.1109/cvpr.2019.01131.

[27] Zhang X E, Ng R, Chen Q F. Single image reflection separation with perceptual losses. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.4786–4794. DOI: 10.1109/cvpr.2018.00503.

[28] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In *Proc. the 14th European Conference on Computer Vision*, Oct. 2016, pp.694–711. DOI: 10.1007/978-3-319-46475-6_43.

[29] Backhaus W G K, Kliegl R, Werner J S. Color Vision: Perspectives from Different Disciplines. De Gruyter, 2011.

[30] Cun X D, Pun C M, Shi C. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *Proc. the 34th AAAI Conference on Artificial Intelligence*, Apr. 2020, p.10680–10687. DOI: 10.1609/aaai.v34i07.6695.

Zi-Nuo Li received his B.S. degree in computer science from Huizhou University, Huizhou, in 2023. He worked as a research assistant at University of Macau, Macao, in 2023. He is currently pursuing his Ph.D. degree with the University of Western Australia, Perth. His current research interests include computational photography, computer vision, and artificial intelligence.

Xu-Hang Chen received his M.Sc. degree in computer and information technology from the University of Pennsylvania, Philadelphia, in 2019. He is currently pursuing his Ph.D. degree with the Department of Computer and Information Science, University of Macau, Macao. His current research interests include computational photography and artificial intelligence.

Shu-Na Guo received her B.S. degree in computer science from Huizhou University, Huizhou, in 2023. She worked as a research assistant at University of Macau, Macao, in 2023. She is currently pursuing her M.Sc. degree with Monash University, Melbourne. Her current research interests include computer vision and artificial intelligence.

Shu-Qiang Wang received his Ph.D. degree in system engineering and engineering management from City University of Hong Kong, Hong Kong, in 2012. He was a research scientist of the Noah's Ark Laboratory, Huawei Technologies, Shenzhen. He held a Post-Doctoral Fellowship at the University of Hong Kong, Hong Kong, from 2013 to 2014. He is currently a professor at the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen. His research interests include machine learning, medical image computing, and optimization theory.

Chi-Man Pun received his Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2002. He was the head of the Department of Computer and Information Science, University of Macau, Macao, from 2014 to 2019, where he is currently a professor and in charge of the Image Processing and Pattern Recognition Laboratory. He has investigated many externally funded research projects as a principal investigator and has authored/coauthored more than 200 refereed papers in many top-tier journals and conferences. He also has two U.S. patents granted. His research interests include image processing and pattern recognition, multimedia information security, forensic and privacy, adversarial machine learning, and AI security.