Frequency-domain Transformer-based Low-light Image Enhancement Network

Yang Zhou
School of Information Engineering
Zhejiang Ocean University
Zhoushan, China
zhouyang@zjou.edu.cn

* Qihong Ye
School of Information Engineering
Zhejiang Ocean University
Zhoushan, China
* Corresponding author: 455657538@qq.com

Hongming Chen
Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province
School of Information Engineering
Zhejiang Ocean University
Zhoushan, China
2022048@zjou.edu.cn

Xiaoshuang Wang
School of Information Engineering
Zhejiang Ocean University
Zhoushan, China
wxshuang@126.com

2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS) | 979-8-3503-2538-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCTIS58954.2023.10213021

***Abstract*—During the process of capturing images under low-light conditions, images often suffer from issues such as low contrast and high noise due to the limitations of illumination. Recently, with the rapid development of transformers, they have shown significant effectiveness in low-light image enhancement tasks. Therefore, we propose a low-light image enhancement network based on the frequency-domain transformer (FDT). This network consists of a main network and a frequency-domain auxiliary network. In the main network, we use the proposed frequency-domain transformer (FDT) to distinguish between high and low-frequency domains and selectively preserve the required high and low-frequency information. In the frequency-domain auxiliary network, the frequency-domain enhancement module (FDEM) is used to extract features and aggregate with the main network, to achieve the purpose of assisting the integration of information from the main network. Extensive experimental results demonstrate that our proposed method outperforms existing methods to a large extent.**

Keywords-frequency-domain; transformer; low-light

I. INTRODUCTION

When capturing images under low-light conditions, the images often suffer from poor illumination, and due to the complex variability of low-light situations, there are significant differences in images captured in different scenes and environments. To address this highly challenging task, many techniques have been proposed for optimization.

Initially, traditional image enhancement methods mainly included histogram equalization, gray-level stretching, and filtering. Methods based on histogram equalization enhance the brightness and contrast of an image by adjusting the pixel grayscale distribution. However, this method often causes excessive enhancement, leading to the appearance of noise in the image. Methods based on the Retinex theory propose an insightful hypothesis that decomposes an image into two components: illumination and reflection. However, the early single-scale Retinex (SSR) and multi-scale Retinex (MSR) methods did not produce ideal visual effects and suffered from problems of excessive enhancement and distortion.

In recent years, with the emergence and continuous development of deep learning, many models have been developed and applied to low-level vision tasks, such as denoising, dehazing, and super-resolution. These methods effectively solve the noise and distortion problems in the image enhancement process by establishing deep neural network models and performing end-to-end learning with a large amount of training data. In 2019, KinD [1] designed a model that provides a mapping function that can flexibly adjust the illumination level according to the user's different needs. The model not only restores details and information in the image by estimating the brightness and color information of each pixel but also reduces noise by smoothing the estimated pixel values. In 2020, Zero-DCE [2] proposed zero-reference deep curve estimation for weak light image enhancement. This method designs a specific curve for the image, which can approximate the pixel-level and high-order curves by iteratively applying itself. Such an image-specific curve can effectively perform mapping within a wide dynamic range. By using these curves to adjust the input image at the pixel level, an enhanced image can be obtained.

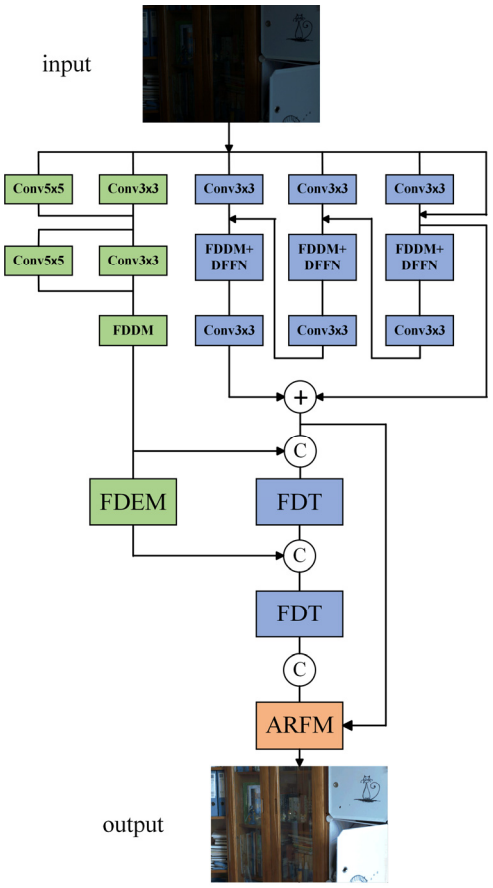
Recently, transformer-based models have also made significant progress in low-light image enhancement. In 2022, 90K [3] proposed a lightweight local-global dual-branch network for low-light image quality enhancement, which converts the image to the raw-RGB domain through an encoder and simulates the ISP process to the sRGB domain through a decoder to achieve image enhancement.

Based on previous work, in this paper, we propose a low-light image enhancement network based on a frequency-domain transformer, aiming to distinguish and process high and low-frequency information in the image through frequency-domain analysis. The contributions of our work are as follows:

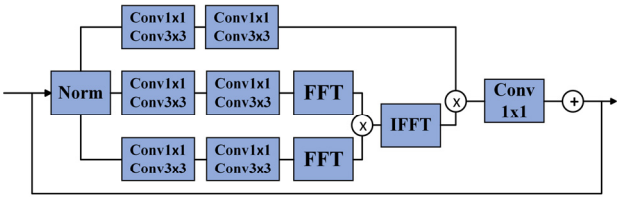
- We innovatively design the network architecture of the frequency-domain transformer (FDT) in the main network, which distinguishes and restores high and low-frequency information to achieve fast transmission of the required frequency information, while effectively improving the network's operating efficiency.
- We effectively combine the multi-scale block and frequency-domain discrimination module (FDDM) in the frequency-domain auxiliary network to correct and enhance important detail information and assist in the transmission of information in the main network.
- We conducted extensive experiments, and the results demonstrate that our proposed network performs well.

II. METHOD

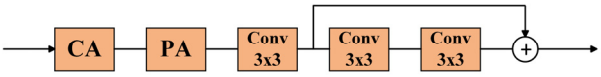
An ideal low-light image enhancement network should output well-exposed images. To achieve this highly challenging task, we developed a low-light image enhancement network based on a frequency-domain transformer (FDTNet). In this section, we first describe the overall architecture of the network, followed by detailed explanations of each module. Figure 1 shows the overall architecture of our network.



(a)FDTNet



(b)FDDM



(c)ARFM

Figure 1. The structure of FDTNet、FDDM and ARFM.

A. *Overall framework*

The overall network consists of three modules: the frequency-domain transformer (FDT), the frequency-domain enhancement module (FDEM), and the attention-based residual fusion module (ARFM). The network takes a low-light image as input, which is processed by the FDT in the main network and the FDEM in the frequency-domain auxiliary network, and finally passed through the ARFM. In the main network, we design a new transformer combination algorithm that uses the transformer's frequency-based self-attention mechanism to estimate the dot-product attention. Furthermore, we use DFFN [7] to determine whether to preserve or enhance high and low-frequency information. In the frequency-domain auxiliary network, we efficiently combine the multi-scale block and frequency-domain discrimination module (FDDM) to enhance and correct important high and low-frequency information. Additionally, we effectively aggregate information between the main network and the frequency-domain auxiliary network. Finally, we further enhance the output using the attention-based residual fusion module (ARFM) to obtain well-exposed images.

B. *Frequency-domain Transformer-based Main Network*

Our main network consists of three FDT modules and several 3×3 convolutional blocks. To avoid losing the original input information, we carefully designed the structure of the FDT module, and for each FDT module, the original information is passed through three times. For the main network, we input a low-light image X with spatial resolution $C\times H\times W$. First, we use a 3×3 convolution and skip connection, followed by layer normalization to normalize each feature of each sample in the network to have a mean of 0 and a variance of 1. Next, we use two sets of 1×1 point-wise convolution and 3×3 depth-wise convolution blocks to obtain features F_q, F_k and F_v. Then, we apply fast Fourier transform (FFT) to estimate the correlation in the frequency domain of the estimated features F_q and F_k by transforming them into frequency domain representations:

$$A = F^{-1}(F(F_q)\overline{F(F_k)}) \tag{1}$$

$F(\cdot)$ denotes the fast Fourier transform (FFT), $F^{-1}(\cdot)$ denotes the inverse FFT, and $\overline{F(\cdot)}$ denotes the conjugate transpose operation.

Next, we estimate the aggregated features through the following approach:

$$V_{att} = L(A)F_v \tag{2}$$

which the layer norm $L(\cdot)$ is used to normalize A.

Finally, we obtain the output features of the frequency domain transformer (FDT):

$$X_{att} = X + Conv1\times1(V_{att}) \tag{3}$$

Through the frequency domain transformer (FDT) based approach, we can effectively separate information into high-frequency and low-frequency components and selectively attend to them to recover low-light images. To determine which high and low-frequency information is useful for our network, we use a DFFN to adaptively determine the frequency information to retain. The DFFN is implemented using the JPEG compression algorithm, where a learnable quantization matrix W is introduced and learned using the inverse method of the JPEG compression to preserve the necessary frequency information.

C. Frequency-domain Auxiliary Network based on FDDM

The auxiliary network consists of two Frequency Domain Enhancement Modules (FDEM). The FDEM module is mainly composed of a 3×3 convolution, a 5×5 convolution, and a Frequency Domain Discrimination Module (FDDM). The auxiliary network takes in a low-light image X with dimensions $C\times H\times W$, and passes it through a convolution block consisting of two 3×3 and two 5×5 convolutions, resulting in four intermediate feature maps. Additionally, this process increases the receptive field and enhances the local detail expression of features. These feature maps are then concatenated and passed through a convolutional feature fusion layer, generating the output feature map. The output is then added to the input X, completing the residual connection. We selectively fuse features from different convolutional layers and improve output quality through residual connections, while enhancing edges and preserving details. Finally, the Frequency Domain Discrimination Module (FDDM) is used to further process the frequency domain information of the image, obtaining more high- frequency and low-frequency semantic information to further improve network performance. The final output of the Frequency-domain Auxiliary Network based on FDDM can be expressed by the following equations:

$$X_{conv3\times3} = Conv_{3\times3}(X) \tag{4}$$

$$X_{conv5\times5} = Conv_{5\times5}(X) \tag{5}$$

$$X_{conv3\times3} = Conv_{3\times3}(Cat(X_{conv3\times3}, X_{conv5\times5})) \tag{6}$$

$$X_{conv5\times5} = Conv_{5\times5}(Cat(X_{conv3\times3}, X_{conv5\times5})) \tag{7}$$

$$X_{out} = FDDM((Cat(X_{conv3\times3}, X_{conv5\times5})) \tag{8}$$

D. Attention-based Residual Fusion Module

After the final convergence of the network, we use attention-based residual fusion modules to further enhance the low-light image enhancement effect. The module consists of channel attention (CA) and pixel attention (PA) modules, as well as a residual module, where the channel attention and pixel attention modules adjust the weights of channels and pixels, respectively, to improve the expressive power and performance of the model. Specifically, the channel attention module includes an adaptive average pooling layer, two convolutional layers, and a sigmoid activation function to generate channel attention weights. Moreover, the pixel attention module aims to learn the inter-channel correlations of the input tensor by down-sampling it with adaptive pooling and generates a scalar weight between 0 and 1 to multiply with the input. The resulting output is influenced by different attention weights for each channel, thus improving the performance of the model. Additionally, the residual module is added to avoid information loss and improve the accuracy of the network. The final output of the attention-based residual fusion module is expressed by the following equations:

$$X_{PA} = PA(x) \tag{9}$$

$$X_{CA} = CA(X_{PA}) \tag{10}$$

$$X_{ARFM} = Res(X_{CA}) \tag{11}$$

III. Loss for FDTNet

Based on the analysis of experimental results, the total loss function used by our low-light image enhancement network is as follows:

$$L_{Net} = \left\| I - I_{gt} \right\|_{L_1} \tag{12}$$

where the enhanced images I are obtained by Net, and the I_{gt} represent the ground truth images.

IV. EXPERIMENTS

A. Dataset

The LOL dataset is created specifically for the challenge of low-light image enhancement. The dataset consists of 500 pairs of images, with 485 pairs used for training and 15 pairs for testing, and each low-light image is paired with a corresponding well-exposed image.

B. Experimental Details

Entire network was trained and tested on the LOL dataset using an Nvidia GTX 2080Ti GPU. The initial learning rate was set to 0.0001, batch size was set to 2, and the total training steps were set to 500,000. Additionally, we used a cosine annealing strategy to adjust the learning rate during training of the network.

Figure 2. Experiments on the LOL dataset.

C. Performance Evaluation

We use two metrics, PSNR and SSIM, to quantitatively compare the performance of the network, where higher values of PSNR and SSIM indicate better quality. We compare the performance with EnGAN [4]、DRBN [5]、Zero-DCE [2]、RetinexNet [6] and KinD [1].

TABLE I. EXPERIMENTAL RESULTS ON LOL DATASET.

Model	PSNR	SSIM
EnGAN[4]	17.48	0.65
DRBN[5]	20.13	0.83
Zero-DCE[2]	14.86	0.54
RetinexNet [6]	16.77	0.56
KinD[1]	20.87	0.80
FDTNet	**20.92**	**0.81**

D. Ablation Experiment

We analyzed the impact of some architecture modules on the final performance of the network. Table II shows that removing the FDEM module at the auxiliary network leads to a decrease in PSNR of 0.75 dB. The absence of the FDEM module in the auxiliary network causes the main network to lose some frequency domain information, resulting in a decrease in PSNR. Moreover, deleting the ARFM module results in a PSNR decrease of 1.12 dB, highlighting the importance of the final residual fusion after aggregation between the main and auxiliary networks. The table below shows the results of our ablation experiments.

TABLE II. RESULTS OF ABLATION EXPERIMENT

Model	W/o FDEM	W/o ARFM	FDTNet
PSNR	20.17	19.8	**20.92**
SSIM	0.79	0.77	**0.81**

V. CONCLUSIONS

We propose a low-light image enhancement network based on frequency-domain transformer, named FDT-Net. Inspired by previous visual network architectures, our network consists of a transformer-based main network and a frequency-domain auxiliary network based on FDDM. Specifically, the main network distinguishes between high and low-frequency domain information of the image and preserves the required high and low-frequency domain information. The frequency-domain auxiliary network uses multi-scale convolutional blocks and frequency-domain discrimination module (FDDM) to enhance the frequency-domain information of the image. The network can ultimately output well-exposed results and significantly outperforms the current techniques. Additionally, there is still room for further improvement of the network, which will be our future work.

ACKNOWLEDGMENT

This work was supported by Basic Scientific Research Fund of Zhejiang Provincial Universities under Grant No. 2021JD004.

REFERENCES

[1] Y. Zhang, J. Zhang, and X. Guo, “Kindling the darkness: A practical low-light image enhancer,” in Proceedings of the 27th ACM international conference on multimedia, pp. 1632-1640, October 2019.

[2] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, “Zero-reference deep curve estimation for low-light image enhancement,” in Proceedings of the IEEE/CVF conference on computer vision andpattern recognition, pp. 1780–1789, 2020.

[3] Z. Cui, K. Li, L. Gu, S. Su, P. Gao, Z. Jiang, Y. Qiao, and T. Harada, “You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction,” in 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24.

[4] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, “Enlightengan: Deep light enhancement without paired supervision,” IEEE transactions on image processing, vol. 30, pp. 2340–2349, 2021.

[5] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, “From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3063–3072, 2020.

[6] C. Wei, W. Wang, W. Yang, and J. Liu, “Deep retinex decomposition for low-light enhancement,” arXiv preprint arXiv:1808.04560 (2018).

[7] L. Kong, J. Dong, M. Li, J. Ge, and J Pan. “Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring, ” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5886-5895, 2023.