mathematics

MDPI

Article

Low-Light Image Enhancement by Combining Transformer and Convolutional Neural Network

Nianzeng Yuan [1], Xingyun Zhao [1], Bangyong Sun [1,2,3,*], Wenjia Han [2], Jiahai Tan [3], Tao Duan [3] and Xiaomei Gao [4]

[1] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
[2] Key Laboratory of Pulp and Paper Science & Technology of Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
[3] State Key Laboratory of Transient Optics and Photonics, Chinese Academy of Sciences, Xi'an 710119, China
[4] Xi'an Mapping and Printing of China National Administration of Coal Geology, Xi'an 710199, China
* Correspondence: sunbangyong@xaut.edu.cn

Abstract: Within low-light imaging environment, the insufficient reflected light from objects often results in unsatisfactory images with degradations of low contrast, noise artifacts, or color distortion. The captured low-light images usually lead to poor visual perception quality for color deficient or normal observers. To address the above problems, we propose an end-to-end low-light image enhancement network by combining transformer and CNN (convolutional neural network) to restore the normal light images. Specifically, the proposed enhancement network is designed into a U-shape structure with several functional fusion blocks. Each fusion block includes a transformer stem and a CNN stem, and those two stems collaborate to accurately extract the local and global features. In this way, the transformer stem is responsible for efficiently learning global semantic information and capturing long-term dependencies, while the CNN stem is good at learning local features and focusing on detailed features. Thus, the proposed enhancement network can accurately capture the comprehensive semantic information of low-light images, which significantly contribute to recover normal light images. The proposed method is compared with the current popular algorithms quantitatively and qualitatively. Subjectively, our method significantly improves the image brightness, suppresses the image noise, and maintains the texture details and color information. For objective metrics such as peak signal-to-noise ratio (PSNR), structural similarity (SSIM), image perceptual similarity (LPIPS), DeltaE, and NIQE, our method improves the optimal values by 1.73 dB, 0.05, 0.043, 0.7939, and 0.6906, respectively, compared with other methods. The experimental results show that our proposed method can effectively solve the problems of underexposure, noise interference, and color inconsistency in micro-optical images, and has certain application value.

Keywords: image processing; deep learning; low-light image enhancement; self-attention mechanism

MSC: 68T07

Citation: Yuan, N.; Zhao, X.; Sun, B.; Han, W.; Tan, J.; Duan, T.; Gao, X. Low-Light Image Enhancement by Combining Transformer and Convolutional Neural Network. *Mathematics* **2023**, *11*, 1657. https://doi.org/10.3390/math11071657

Academic Editor: Jakub Nalepa

Received: 1 March 2023
Revised: 21 March 2023
Accepted: 27 March 2023
Published: 30 March 2023

1. Introduction

Images captured in low-light environment usually result in a series of image degradation problems, such as low contrast, poor visibility, noise pollution, or color distortion. These degradation problems prominently reduce the image perception quality and potentially hamper the subsequent image processing tasks, e.g., image segmentation [1], target recognition [2], video surveillance [3] and anomaly detection [4]. Although the image brightness can be improved partly by extending exposure time, undesired noise is likely to be amplified on the optical sensors. Therefore, it is significant to develop effective enhancement methods for low-light images to raise their perception quality as taken in normal lighting conditions.

In the past decades, various low-light image enhancement methods have been proposed based on specific mathematical models, such as the histogram equalization method

and Retinex method, etc. In reality, various artifacts often appear in those enhanced images, e.g., too much noise [5], blurring, or false colors. Recently, due to the strong representation capability of CNN, many CNN-based models have been proposed to enhance the low-light image. However, most CNN-based approaches have difficulties in learning global semantic information, resulting in that some unsatisfied artifacts such as uneven brightness or color distortion usually exist in those models. As shown in Figure 1, the images (b) and (c) are the enhanced results of the histogram equalization method and the CNN-based method, respectively. It can be clearly seen that these two enhanced images are not satisfactory. The enhancement results of LIME method obviously have the problems of chromatic aberration, low brightness, and slight noise. The enhancement results of RetinexNet method also have obvious problems of chromatic aberration and low brightness, and the whole image is accompanied by strong noise. While in Figure 1d of the image enhanced by our proposed method, it looks natural, with no false colors or noise. Generally, by referring the above works, we observe that two major challenges still exist in the task of low-light image enhancement: (I) How to adaptively enhance the image brightness of different exposure areas at uneven illumination condition; (II) how to suppress the image noise while maintaining the consistency of color textures.

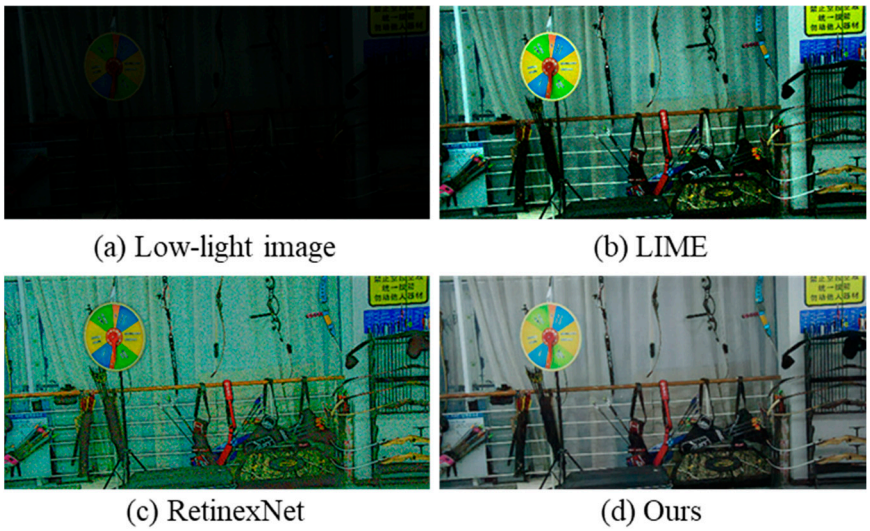

Figure 1. Results from various methods on an image from the LOL dataset. (**a**) The input images; (**b**) the results of LIME [6]; (**c**) the results of RetinexNet [7]; (**d**) the results of ours. It is obvious that the existing method has serious noise and does not enhance the image brightness correctly.

To address the challenges above, it is useful to obtain accurate local and global features of low-light images. The low-light image enhancement task involves not only the refinement of image brightness and color recovery, but also the suppression of image noise. From the image denoising task, we can determine the noise location by obtaining the global correlation of the image, thus learning that it is crucial to obtain the global features of the image. In the image recovery process, we need to use local features to retain important detail information of the image. Generally, the global features and local features are the key for adaptively enhancing image brightness at uneven illumination condition. For these reasons, this paper will discuss how to fully acquire global and local features of low-light images.

Many CNN-based enhancing models learn the global features by increasing the receptive field of the convolution kernel. The receptive field increasing is usually implemented by means of adding more layers of the CNN network or applying down-sampling multiple times. In reality, adding layers inevitably introduces more network parameters, while employing multiple down-sampling frequently results in the loss of image details. Inspired by the great success of transformer [8] in the field of natural language processing (NLP) and computer vision (CV), we apply the transformer for global feature extraction to com-

pensate the shortcomings of CNN. Basically, the transformer is quite different from the convolution operation of CNN during feature extraction. The CNN convolution focuses on extracting the local relationship between the center point and its surrounding pixels, while the transformer utilizes multi-head self-attention and forward propagation network, which is beneficial for extracting long-term dependencies.

Since transformer and CNN have their own advantages for extracting features, in this paper, we propose an end-to-end enhancement network combining transformer and CNN to address the challenges above. Essentially, the goal of our proposed enhancement network is to effectively extract image local and global features, which is significant to boost image brightness for different image regions and suppress noise or chromatic aberrations. Within the proposed enhancement network, we creatively build a fusion block by combining the advantages of transformer and CNN to obtain the features simultaneously. The CNN stem mainly extracts the local detailed feature information of the image by utilizing convolutions, while the transformer stem focuses on learning to global semantic information by self-attention mechanism.

As shown in Figure 2, the structure of the proposed enhance network is similar to a U-Net consisting of a group of symmetric fusion blocks with skip connections. On the left, the fusion blocks combined with down-sampling are utilized to extract the semantic features from the low-light image. On the right, the fusion blocks combined with up-sampling are used to recover the normal light image from semantic features. In parallel, skip connections are utilized to minimize the loss of spatial information due to down-sampling. In summary, contributions of this paper are described as follows:

1. We propose an end-to-end low-light image enhancement network by combining the transformer and CNN. Both the local and the global features are accurately learned for light enhancement within this network, where the CNN effectively extracts local features and the transformer precisely learns the long-range dependencies.
2. We creatively applied the transformer model to the low-light image enhancement task, and built a U-shaped low-light network with fusion blocks including transformer and CNN.
3. Our proposed method is evaluated on the LOL dataset, and the experimental results demonstrate that the proposed network outperforms the other state-of-the-art (SOTA) models in qualitative and quantitative comparisons.

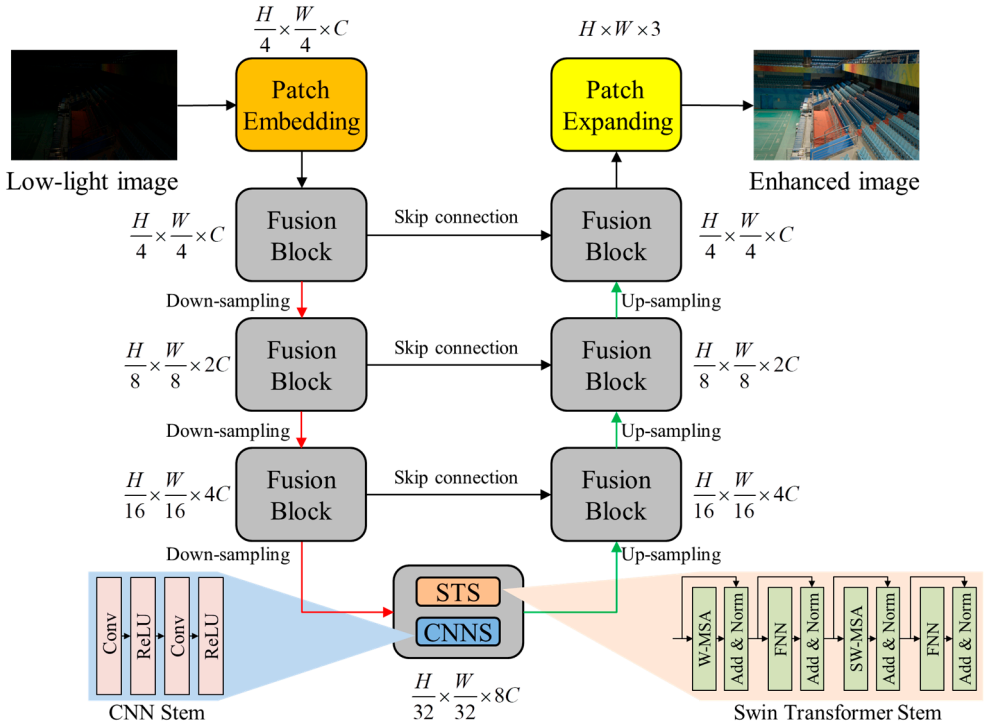

Figure 2. Overview of our network structure.

2. Related Work

Various works on low-light image enhancement can be described in detail below.

2.1. Conventional Low-Light Image Enhancement Methods

The conventional low-light image enhancement methods are roughly divided into histogram equalization-based methods and illumination-based methods. The histogram equalization [9] method directly adjusts the spatial distribution of the image brightness histogram to improve the image contrast. Ibrahim et al. [10] and Wang et al. [11] modified the histogram equalization method by adding a brightness threshold to improve the image brightness. Overall, the mentioned histogram methods are simple to operate, but it is not possible to ensure that all areas of the image can be improved without taking into account the spatial information of the image manually. Chen et al. [12] proposed a local histogram equalization method to adaptively solve the problem of image spatial distribution, and divided the image into blocks for brightness enhancement. This method performs fast and can enhance the image details well except for the block effects.

Illumination-based methods are mainly based on Retinex theory, which decomposes images into illumination components and reflection components. Jobson et al. [13] proposed the Multi-Scale Retinex (MSR) algorithm, which utilized Gaussian filtering at different scales to perform illumination decomposition on images, and then the brightness enhancement is implemented on the illumination components. Wang et al. [14] added a logarithmic bilateral conversion based on the Retinex theory to make the obtained illumination components closer to natural colors. Li et al. [15] proposed a more robust Retinex algorithm to enhance low-light images by adding noise processing. These methods are effective in illumination adjustment and noise elimination. However, the parameters in the model need to be set artificially, thus it is difficult to adaptively handle the variety of images or some images with strong noise.

2.2. Intelligent Low-Light Image Enhancement Methods

In recent years, along with the rapid development of artificial intelligence theory, learning-based low-light image enhancement methods have been consistently introduced. Lore et al. [16] proposed a deep auto-encoder network named LLNet for contrast enhancement and noise removal. Wei et al. [7] proposed a neural network named RetinexNet, which employed two sub-networks to decompose low-light images and adjust the illumination components. Zhang et al. [17] proposed the KinD network to optimize the [7] method by adding a restoration network to denoise the reflection components. Chen et al. [18] developed the SID network to directly perform low-light image enhancement on raw sensor data. Wang et al. [19] construed an illumination estimation network named DeepUPE, which adopted an end-to-end form to learn the image-light mapping relationship to predict the smooth light mapping. Jiang et al. [20] proposed the unsupervised adversarial network named EnlightenGAN, which utilized a global-local discriminator and a self-regularized attention mechanism to process the synthetic as well as the real-world images. Ma et al. [21] propose a self-correcting shared-weight illumination learning module for low-light image enhancement, which substantially reduces the large inference cost under the cascade mechanism and greatly reduces the computational effort of the network model. Although, the method based on deep learning can offset the previous methods to a certain extent and achieve better visual results for most low-light images. However, the above methods still do not produce satisfactory visual quality for low-light images.

3. Proposed Method

We proposed a low-light image enhancement network by fusing transformer with CNN to reconstruct a high-quality noise-free color image. The enhancement of the low-light image can be defined as:

$$I_{\mathrm{h}} = F(I_i, \sigma) \tag{1}$$

where I_i is a low-light image input, I_h stands for the enhanced image, F means the low-light image enhancement network, and σ represents the parameters involved.

3.1. Network Framework

As shown in Figure 2, our proposed end-to-end enhancement network takes a low-light image as input and obtains an enhanced high-quality color image. The network is designed into U-shape which consists of a patch embedding module, a cascade of fusion blocks, and a patch expanding module. First, the patch embedding module is utilized to convert the input image into the patch embedding codes. And then, the patch embedding codes are sent to the fusion blocks arranged symmetrically with skip connection. Each fusion block with a Swin Transformer stem and a CNN stem inside will further extract the deep features from the previous one. Meanwhile, the down-sampling and up-sampling are utilized to adjust the resolution of calculated features from each fusion block. Finally, the patch expanding module is used to modify high-resolution feature dimensions by feature mapping and to output enhanced results.

3.2. Fusion Block

As shown in the bottom of Figure 2, each fusion block consists of a Swin Transformer stem and a CNN stem connected in parallel. The transformer stem is utilized to learn global semantic information and capture long-term dependencies, while the CNN stem is capable of efficiently extracting local features and emphasizing detailed features such as image color and texture. We describe these two stems in detail as below.

3.2.1. Swin Transformer Stem

As shown in the right bottom of Figure 2, the Swin Transformer stem includes a Normalization layer (Norm), a Windows-based Multi-head Self-Attention layer (W-MSA), a Shifted Window based Multi-head Self-Attention layer (SW-MSA), and a Feed-forward Neural Network layer (FNN). The Norm layer mainly serves to perform batch regularization and normalize the input data, for the purpose of ensuring the regularity of the data distribution in the input layer. In order to reduce the self-attention calculation complexity, the input image is usually divided into different window regions, and the self-attention calculations are performed in the smaller window regions. At the same time, the contents of the window are changed by circular displacement to ensure the interaction of global information. W-MSA and SW-MSA represent self-attention calculation within different window regions, while the self-attention [8] calculation can be defined as:

$$Attention(Q, K, V) = SoftMax(\frac{QK^{\mathrm{T}}}{\sqrt{d}} + B)V \quad (2)$$

where Q means the query matrix, K is the key matrix, V represents the values matrix, d stands for the dimension of the query matrix or key matrix, and B is the position matrix.

3.2.2. CNN Stem

As shown in the left bottom of Figure 2, the CNN stem consists of several convolutional layers and activation functions. The relationship between the input pixel and its neighbors can be obtained through the convolution operation. The neural network efficiently extracts the local features of the image and accurately preserves the details of the image by the inherent bias of the convolution operation.

As shown in the Figure 3, the global features obtained by Swin Transformer stem and the local features by CNN stem are concatenated in the channel dimension, and then the number of channel dimensions is adjusted by using the fully connected layer, and the ReLU activation function increases its nonlinearity.

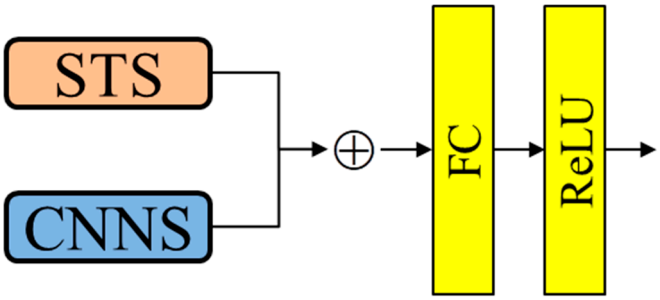

Figure 3. The fusion structure of the output features of Swin Transformer stem and CNN stem.

3.3. Loss Function

A combined loss function was constructed to accurately recover the image brightness and colors with low noises. This loss function mainly consists of L_1 loss function [22], structural similarity (SSIM) loss function [23], and perceptual loss function [24] as follows:

$$L_{\text{total}} = (1 - \lambda_{\text{s}} - \lambda_{\text{p}})L_1 + \lambda_{\text{s}}L_{\text{ssim}} + \lambda_{\text{p}}L_{\text{perc}} \tag{3}$$

where L_1 represents the pixel-level parametric loss, L_{ssim} stands for the structural similarity loss, L_{perc} means the perceptual loss, while λ_{s} and λ_{p} are two adjusting coefficients.

3.3.1. Pixel-Level Parametric Loss

The L_1 loss is better to reduce the difference between the predicted image and the real image by calculating the average distance pixel by pixel. Therefore, we adopted L_1 loss to optimize our model, as follows:

$$L_1 = \sqrt{\| I_{gt} - I_h \|^2 + \ell}(\ell = 10^{-6}) \tag{4}$$

where I_{gt} represents the real image, I_h stands for the predicted image, ℓ is a non-zero constant.

3.3.2. Structural Similarity Loss

The structural similarity loss function measured the structural loss of real and predicted images in three aspects: brightness, contrast, and image structure, which contributes to recovering the structure and local details of images, as follows:

$$L_{ssim} = 1 - \frac{1}{N}\sum_{img}\left(\frac{2\mu_x\mu_y + C_1}{{\mu_x}^2 + {\mu_y}^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \tag{5}$$

where μ and σ^2 represent the image mean and variance, respectively, C_1, C_2 are two constants.

3.3.3. Perceptual Loss

The perceptual loss mainly constrains the perceptual difference between the real image and the predicted image, for maintaining the image perception, veracity of details, and semantic fidelity, as follows:

$$L_{perc}(I_{gt}, I_h) = \frac{1}{C_j H_j W_j} \| \varphi(I_{gt}) - \varphi(I_h) \|_1 \tag{6}$$

where I_{gt} represents the real image, I_h stands for the predicted image, H_j and W_j are the height and width of the j layer feature map, respectively, C_j is the channel, $\varphi()$ is the feature maps obtained in the pre-trained VGG16 model.

4. Experiments

In this section, we perform experiments on a public low-light image dataset to evaluate our proposed method. First, we introduced the dataset, implementation details, and evaluation criteria. Then, we compared the proposed method with the state-of-the-art methods in a public low-light image test set. Finally, we performed a series of ablation experiments to verify the effectiveness of the proposed fusion block in the network.

4.1. Experimental Setup

4.1.1. Datasets

We choose LOL dataset provided by Wei et al. [7] to evaluate our method. The LOL dataset includes 789 low-light image pairs captured in real extreme dark environments by varying the exposure time of the camera to obtain image data pairs with noise, where the first 689 pairs are utilized to train the network model, and the last 100 pairs are used for testing. In addition, to ensure the generalization performance of the proposed network model, we add 900 synthetic low-light image pairs during network training.

4.1.2. Implementation Details

We implemented our network in Pytorch [25] and trained it for 1000 epochs on an Nvidia GTX2080 GPU with 64 GB of memory. During training, the batch size is set 4 and the learning rate is 0.0001, which is optimized by the ADAM optimizer.

4.1.3. Evaluation Criteria

To quantitatively evaluate the low-light image enhancement performance of different methods, we have chosen to evaluate with and without reference evaluation metrics, respectively. The reference evaluation metrics include PSNR [26], SSIM [23], LPIPS [27], DeltaE [28], and DeltaE2000 [29], and the no-reference evaluation metrics are NIQE [30]. The PSNR evaluation metric focuses on measuring the fidelity similarity between images, with higher values indicating more similarity between the two images. The SSIM evaluation metric measures the similarity between two images in terms of brightness, contrast, and structure, and higher values indicate more similarity between the two images. The LPIPS evaluation metric mainly evaluates the perceptual distance between image features, and lower values indicate more similarity between the two images. The DeltaE and DeltaE2000 evaluation metrics mainly calculate the color difference between image pixels, and the lower the value means the smaller the color difference between the two images. The NIQE is a non-reference evaluation metric, which mainly measures the difference in multivariate distribution of images and is closer to the human vision system.

4.2. Comparison with Other Methods

To verify the effectiveness of our proposed method, we compare our proposed method with several state-of-the-art (SOTA) methods. The selected SOTA methods include LIME [6], MBLLEN [31], RetinexNet [7], KinD [17], GLADNet [32], SIE [33], and Zero-DCE [34]. To ensure the fairness of the experiments, all methods are tested in the same experimental environment.

4.2.1. Qualitative Comparison

We first compare the visual performance of our enhanced low-light images with contrasting methods on the LOL test set. As shown in Figure 4, we selected four representative images on the LOL test set [13] for qualitative comparison, and the enhancement results of our proposed method are visually significantly better than the other methods. The visualization results of RetinexNet and Zero-DCE methods were the worst, and Zero-DCE method could not fully enhance the image brightness, and the enhancement results also had insufficient brightness in some areas. RetinexNet method was better than Zero-DCE method in image brightness enhancement, but the enhanced image amplified the noise, which seriously affected the image quality. The GLADNet, SIE, and MBLLEN methods

have significantly improved the enhancement effect compared to the previous two methods, and all of them can correctly improve the brightness of the image without noise amplification. However, for detail restoration, the GLADNet, SIE, and MBLLEN methods all suffer from color distortion, and the GLADNet method has an overall yellowish enhancement result, making the image look unnatural overall. Although the enhancement results of the SIE and MBLLEN methods are better than the former in terms of image color, however, the image color is oversaturated and appears to be inconsistent with the reference image color. In LIME and KinD, the enhanced image is a bit brighter, while some slight noise and color distortion remain. Compared with these SOTA methods, our proposed method demonstrates a better restoration of images acquired in extreme darkness, with two major improvements as follows. First, our method is able to suppress noise better and recover more details with texture information. Second, our method can correctly and adaptively enhance the image brightness with good generalization. The reconstructed image demonstrates high quality in terms of correct natural color and high fidelity. In summary, our proposed method not only accurately enhances the brightness of most image areas even captured at uneven illumination conditions, but also achieves excellent performance in image details reconstruction and noise suppression.

In addition, we also conducted subjective evaluation on five public datasets of MEF [35], LIME [6], DICM [36], VV [37], and NPE [14], and selected representative image visualization results on each dataset, as shown in Figure 5. The main difference between these five datasets and the LOL dataset is that the original low-light images have a certain brightness and are not taken in extreme darkness, with the main drawback being the underexposure of local areas of the images. From Figure 5, we could find that the MBLLEN and GLADNet methods have some shortcomings in enhancing the image brightness, and the enhancement results of both methods still have underexposed areas, but basically they can recover the image color correctly and suppress the image noise well. The RetinexNet and SIE methods are better than the previous two methods in enhancing the image brightness, but they do not suppress the noise well and do not repair the color features correctly, resulting in the overall image with a serious color distortion. The enhancement results of KinD, LIME, and Zeroc-DCE methods are better than the above four methods, and the enhanced images do not have extremely serious defects in brightness enhancement, noise suppression, and insurance color consistency problems. The enhancement effect of each method has a different focus, and there are also small defects in different aspects. The defects of each method are described as follows. When the KinD method is used to enhance the images of trees, grasses, ponds, etc., there is color distortion in the enhancement results, which shows that the generalization ability of the KinD model has some room for improvement, as shown in Figure 5 of the enhancement effect of KinD. The enhancement effect of the LIME method is seen to have achieved better results in terms of image brightness and color, but when the image is enlarged, there is a slight layer of noise in the enhanced image as a whole, thus showing that the LIME method has certain disadvantages in noise suppression. The Zero-DCE method achieves better results compared to the above comparison methods in all aspects of image enhancement, but there is a slight overexposure problem, as shown in Figure 5. Compared with the above algorithms, our proposed method achieves better results in all aspects.

Figure 4. Visual comparison with existing methods. (The Chinese characters in first row mean "the membership is encouraged").

In summary, the effect of our proposed method achieves better visual results both in the LOL dataset and in the MEF, LIME, DICM, VV, and NPE public datasets, demonstrating the advantages of our proposed method in image brightness enhancement, noise suppression, color recovery, and preservation of image structure texture.

Figure 5. Visual comparison with existing methods.

4.2.2. Quantitative Comparison

In this paper, the objective metrics PSNR, SSIM, LPIPS, DeltaE, DeltaE2000, and NIQE are selected for evaluation. As shown in Table 1, our proposed method achieves the best results in all three metrics. Compared with the optimal values of other methods, the improvements of PSNR, SSIM, LPIPS, DeltaE, and NIQE are 1.73 dB, 0.05, 0.043, 0.7939, and 0.6906 respectively. Overall, the quantitative as well as the qualitative experiment results show that our proposed method outperforms all those compared SOTA methods.

Table 1. Objective evaluation indicators for different methods.

Method	MBLLEN	RetinexNet	KinD	GLADNet	LIME	SIE	Zero-DCE	Ours
PSNR (dB)	17.73	16.46	19.95	19.50	17.25	17.13	18.53	**21.68**
SSIM	0.706	0.494	0.813	0.756	0.562	0.731	0.659	**0.863**
LPIPS	0.250	0.441	0.138	0.240	0.283	0.259	0.234	**0.095**
DeltaE	18.0333	16.0340	10.8781	11.7905	10.1873	13.4434	15.3983	**10.0842**
DeltaE2000	23.0806	19.7853	15.8863	16.1066	**12.4602**	18.4470	21.6841	**14.6195**
NIQE	5.3745	7.0342	4.8470	4.7715	5.2755	5.4065	4.9446	**4.0809**

4.3. Ablation Study

The ablation experiments are performed on the LOL dataset to verify the effectiveness of the fusion blocks proposed in this paper, with the optimal network structure obtained. The quantitative results are shown in Table 2, and Figure 6 demonstrates the visual comparison results of different network structures within our method. It is clear that our proposed method achieves the best results in both objective and subjective metrics. Details of the ablation experiment are described below.

Table 2. Objective evaluation indicators for different methods.

Backbone	Swin Transformer Stem	CNN Stem	ViT Transformer Stem	Fusion Blocks	PSNR (dB)	SSIM	LPIPS	NIQE
√	√				20.30	0.854	0.107	**4.0778**
√		√	√	7	18.51	0.713	0.240	4.7611
√	√	√		5	20.24	0.851	0.115	4.2105
√	√	√		7	**21.68**	**0.863**	**0.095**	**4.0809**

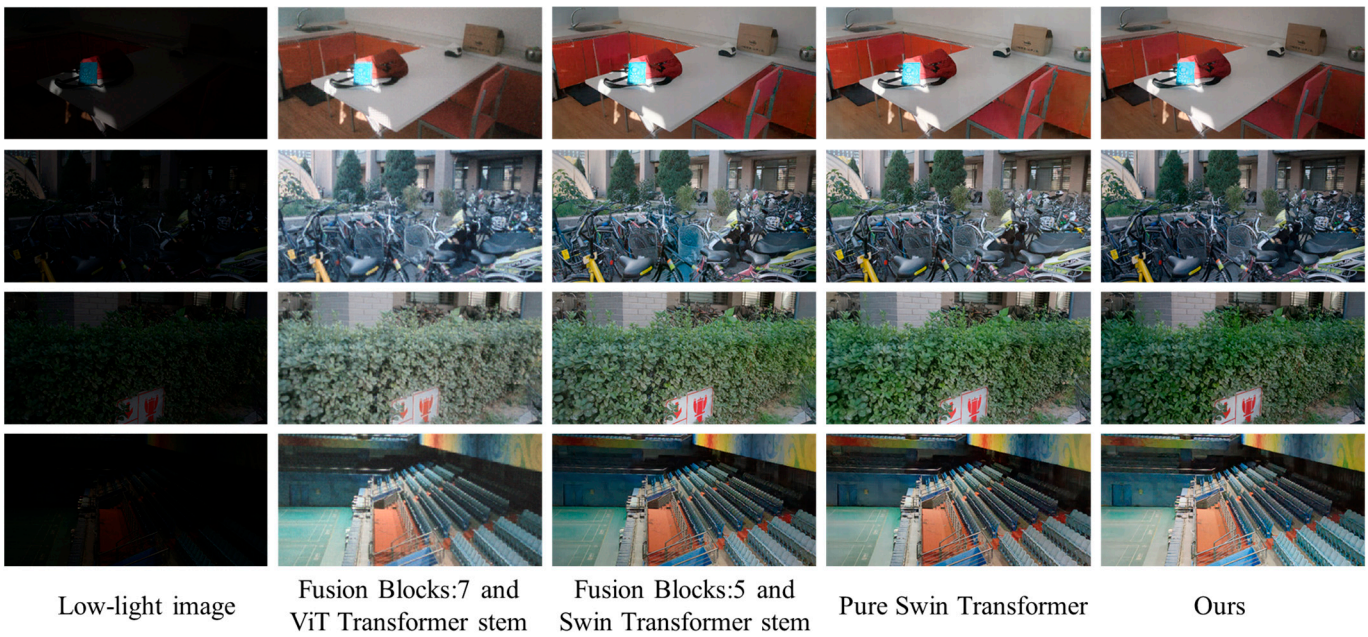

Figure 6. Visual comparison with ablation experiment.

4.3.1. Effectiveness of Fusion Block

To demonstrate the effectiveness of our proposed fusion block, we compare it with the pure Swin Transformer model. To ensure fairness of the comparison, we used the same network depth as well as the experimental parameter settings, and only reduced the CNN stem. As shown in Figure 6 and Table 2, our proposed method increases 1.38 dB, 0.009, and 0.012 in PSNR, SSIM, and LPIPS objective metrics, respectively. Thus, compared with the

pure Swin Transformer model, the experiment result fully demonstrates the effectiveness of our proposed fusion block.

4.3.2. Effectiveness of Different Network Structures

As shown in Figure 6 and Table 2, we have also performed ablation experiments on the performance of different network structures. First, the selection of transformer stem is tested and analyzed in PSNR, SSIM, and LPIPS. We select two frequently used transformers for comparison which are ViT Transformer stem [38] and Swin Transformer stem [39]. It can be seen that the network with Swin Transformer stem performs better than that with ViT Transformer stem. Then, different numbers of fusion blocks are set within the proposed network. As shown in Table 2, all three objective metrics improve significantly when the number of fusion blocks increases from 5 to 7.

5. Conclusions

In this paper, we aim to achieve the task of adaptive brightness enhancement and noise suppression for low-light images. For this goal, we propose an end-to-end low-light image enhancement network by combining transformer and CNN to achieve the reconstruction of luminance information of low-light images. The established network model takes advantage of the feature extraction of transformer mechanism to construct long dependencies between features and extract image semantic features more fully and extensively. Meanwhile, the CNN branch uses convolutional operations, which can acquire local features in more detail and fully preserve image texture and color features, possessing better image recovery effects. Finally, we compare subjectively and objectively the processing effect of this method with the current popular algorithms for low-light images through a large number of experiments, and the results show that this method achieves a better enhancement effect and basically solves the problems of brightness enhancement, noise suppression, and color detail recovery of low-light images, but there is still some room for improvement. In the future work, we will further improve the model in this paper to make it have better generalization performance.

Author Contributions: N.Y. and X.Z. worked on conceptualization, methodology, software and writing—original draft preparation; B.S., W.H., J.T., T.D. and X.G. worked on validation and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62076199, in part by the Key R&D project of Shaanxi Province under Grant 2022ZDLGY01-03, in part by the Open Research Fund of State Key Laboratory of Transient Optics and Photonics, Chinese Academy of Sciences under Grant SKLST202214 and Grant SKLST202005, and in part by the Foundation of Key Laboratory of Pulp and Paper Science and Technology of Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences) under Grant KF202118; and in part by Xi'an science and technology research plan (No. 22GXFW0088).

Data Availability Statement: Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

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

References

1. Li, H.; Xiong, P.; Fan, H. DFANet: Deep feature aggregation for realtime semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9514–9523.
2. Maji, S.; Bourdev, L.D.; Malik, J. Action recognition from a distributed representation of pose and appearance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3177–3184.
3. Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; Saenko, K. Translating videos to natural language using deep recurrent neural networks. *arXiv* **2017**, arXiv:1412.4729.
4. Zhao, Z.; Sun, B.Y. Hyperspectral anomaly detection via memory-augmented autoencoders. *CAAI Trans. Intell. Technol.* **2022**, 1–14. [CrossRef]

5. Zhang, Q.; Xiao, J.Y.; Tian, C.W.; Lin, J.C.; Zhang, S.C. A robust deformed convolutional neural network (CNN) for image denoising. *CAAI Trans. Intell. Technol.* **2022**, 1–12. [CrossRef]
6. Guo, X.J.; Li, Y.; Ling, H.B. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.* **2016**, *26*, 982–993. [CrossRef]
7. Wei, C.; Wang, W.J.; Yang, W.H.; Liu, J.Y. Deep retinex decomposition for low-light enhancement. In Proceedings of the 29th British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018; pp. 1–12.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
9. Vyas, A.; Yu, S.; Paik, J. Fundamentals of digital image processing. In *Signals and Communication Technology*; Prentice-Hall: Upper Saddle River, NJ, USA, 2018; pp. 3–11.
10. Ibrahim, H.; Kong, N.S.P. Brightness preserving dynamic histogram equalization for Image contrast enhancement. *IEEE Trans. Consum. Electron.* **2008**, *53*, 1752–1758. [CrossRef]
11. Wang, C.; Ye, Z. Brightness preserving histogram equalization with maximum entropy: A variational perspective. *IEEE Trans. Consum. Electron.* **2005**, *51*, 1326–1334. [CrossRef]
12. Chen, S.D.; Ramli, A.R. Minimum mean brightness error Bi-histogram equalization in contrast enhancement. *IEEE Trans. Consum. Electron.* **2003**, *49*, 1310–1319. [CrossRef]
13. Jobson, D.J.; Rahman, Z.; Woodell, G.A. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc.* **1997**, *6*, 965–976. [CrossRef] [PubMed]
14. Wang, S.H.; Zheng, J.; Hu, H.M.; Li, B. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Trans. Image Process.* **2013**, *22*, 3538–3548. [CrossRef]
15. Li, M.D.; Liu, J.Y.; Yang, W.H.; Sun, X.Y.; Guo, Z.M. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Trans. Image Process.* **2018**, *27*, 2828–2841. [CrossRef]
16. Lore, K.G.; Akintayo, A.; Sarkar, S. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognit.* **2017**, *61*, 650–662. [CrossRef]
17. Zhang, Y.H.; Zhang, J.W.; Guo, X.J. Kindling the darkness: A practical low-light image enhancer. In Proceedings of the 27th ACM International Conference on Multimedia(ICM), Nice, France, 21–25 October 2019; pp. 1632–1640.
18. Chen, C.; Chen, Q.F.; Xu, J.; Koltun, V. Learning to see in the dark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3291–3300.
19. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6842–6850.
20. Jiang, Y.F.; Gong, X.Y.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.H.; Yang, J.C.; Zhou, P.; Wang, Z.Y. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. [CrossRef]
21. Ma, L.; Ma, T.Y.; Liu, R.S.; Fan, X.; Luo, Z.Y. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
22. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2599–2613. [CrossRef]
23. Wang, Z.; Bovik, A.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
24. Johnson, J.; Alahi, A.; Li, F.-F. Perceptual losses for real-time style transfer and super-resolution. *arXiv* **2016**, arXiv:1603.08155.
25. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.
26. Brauers, J.; Aach, T. A color filter array based multispectral camera. In Proceedings of the Workshop Farbbildverarbeitung, Ilmenau, Germany, 5–6 October 2006; Volume 12, pp. 1–11.
27. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
28. Webster, M.A.; Kay, P. Color categories and color appearance. *Cognition* **2012**, *122*, 375–392. [CrossRef] [PubMed]
29. Luo, M.R.; Gui, G.; Rigg, B. The development of the cie 2000 colour difference formula: Ciede2000. *Color Res. Appl.* **2001**, *26*, 340–350. [CrossRef]
30. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "Completely Blind" image quality analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [CrossRef]
31. Lv, F.F.; Lu, F.; Wu, J.H.; Lim, C.S. MBLLEN: Low-light image/video enhancement using CNNs. In Proceedings of the 29th British Machine Vision Conference(BMVC), Newcastle, UK, 3–6 September 2018; pp. 1–13.
32. Wang, W.J.; Wei, C.; Yang, W.H.; Liu, G.Y. GLADNet: Low-light enhancement network with global awareness. In Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition(FG), Xi'an, China, 15–19 May 2018; pp. 751–755.
33. Zhang, Y.; Di, X.G.; Zhang, B.; Wang, C. Self-supervised image enhancement network: Training with low light images only. *arXiv* **2020**, arXiv:2002.11300.

34. Guo, C.; Li, C.Y.; Guo, J.C. Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1777–1786.
35. Ma, K.; Zeng, K.; Wang, Z. Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process* **2015**, *24*, 3345–3356. [CrossRef] [PubMed]
36. Lee, C.; Lee, C.; Kim, C. Contrast enhancement based on layered difference representation of 2D histograms. *IEEE Trans. Image Process.* **2013**, *22*, 5372–5384. [CrossRef] [PubMed]
37. Vonikakis, V.; Andreadis, I.; Gasteratos, A. Fast centre-surround contrast modification. *IET Image Process.* **2008**, *2*, 19–34. [CrossRef]
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.H.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.
39. Cao, H.; Wang, Y.Y.; Chen, J.; Jiang, D.S.; Zhang, X.P.; Tian, Q.; Wang, M.N. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv* **2021**, arXiv:2105.05537.