Low-Light Image Enhancement based on Zero-DCE and Structural Similarity Loss

Qiyao Li[1] Zhequan Li[2,*] Haoyang Wang[3]

[1] Detroit Green Technology Institute, Hubei University of Technology, 430068, No.28 Nanli Road, Wuhan,Hubei, China
[2]College of Biomedical Engineering and Instrument Science, Zhejiang University, 310027, No.38 Zheda Road, Hangzhou, Zhejiang, China
[3]College of International Education, Shanghai Jian Qiao University, 201306, No.1111 Huchenghuan Road, Shanghai, China
*Corresponding author: lizhequan@zju.edu.cn

Abstract. The computer vision community has become increasingly interested in Low-Light Image Enhancement (LLIE), which tries to transform low-light photos into typically exposed images. The convolutional neural network has advanced quickly, and this has helped the deep learning-based LLIE approaches make a breakthrough in accuracy and visual effects. However, some challenges still remain, especially when dealing with noise from the black color blocks and halo near the boundary of the bright area. In this study, we provide a low-light picture enhancing technique based on the Unet3+ to overcome these problems. Specifically, we first transform DCE-Net in Zero-DCE to Unet3+, which enhances the network's fitting ability. Then, we introduce a denoising module and an SSIM loss, which can improve the qualitative and quantitative metrics of the network. Numerous tests support the effectiveness of our suggested approach, where the normal exposure images produced have a stable brightness and are suitable for a range of scenes.

Keywords: LLIE; UNet3+; image denoising; SSIM loss

1 Introduction

The goal of Low-Light Image Enhancement (LLIE), a crucial basic computer vision task, is to transform low light images into normally exposed images. People frequently generate low light images in daily lives, which may be due to users being unfamiliar with their camera equipment leading to incorrect exposure time setting, or the object is in a backlit environment. Low-light images are a challenge for computer vision. In face verification tasks, for example, the dark environment will produce a dark face which may lead to a decrease recognition accuracy. At the same time, with the widespread application of computer vision in daily life, more and more scenes require that the flash cannot be used during shooting, such as pedestrian detection at night. In this case, turning on the flashing lights may cause dizziness to pedestrians, thereby increasing the risk of traffic accidents. In light of this, there is a great deal of

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_96

interest in research regarding the creation of efficient LLIE algorithms, which also has significant theoretical value and promising application opportunities.

Numerous deep learning-based solutions to the issue of LLIE task have surfaced most recently. By utilizing deep learning techniques, LLNet [1] initially addresses the low-light picture enhancing problem. Retinex-Net [2] makes the assumption that reflectance and illumination can be separated out of an image. Zero-DCE [3] makes use of zero-shot learning, which formulates light enhancement as the endeavor of picture-specific curve estimate and addresses the issue of low-light image improvement by skillfully creating loss functions. Using unsupervised learning, EnlightenGAN[4] converts the LLIE task into an image generation task.

However, some common challenges remain in the research of LLIE. The first difficulty is noise, as seen in Figure 1. The noise that can be ignored in the original image is amplified when the network is enhanced, especially when there are large black color blocks in the original image. The second challenge is halo. When a high-contrast image is enhanced by a network, contrasting color patches are enhanced at the same time, and brighter color patches are enlarged into halos.

(a) Failure enhancement with noise

(b) Failure enhancement with halo

Fig. 1. Visualization of failure low-light image enhancement cases (Photo/Picture credit: Original)

This paper suggests a deep learning-based network for enhancing low-light pictures to address these problems. This paper transforms DCE-Net in Zero-DCE to Unet3+[5], which enhances the network's fitting ability. Use SSIM loss [6] and increase its weight, which can improve the qualitative and quantitative metrics of the network. In addition, try to introduce a denoising module to solve the noise problem, which improves the quantitative metrics. The network made up of these methods generates images with normal exposure that are suitable for a variety of scenes and have stable brightness.

In conclusion, this paper proposes a deep learning-based approach to deal with LLIE task and reports promising outcomes. This paper expands the Unet3+ network, SSIM loss, and denoising modules based on Zero-DCE. The network improves the accuracy of qualitative and quantitative metrics and it generates better augmented pictures. The research findings presented in this paper are crucial to the

implementation and advancement of LLIE technology and serve as a helpful guide for related study and real-world application in the area of computer vision.

2 Method

Figure 2 depicts the overall procedure. The backbone network is an adapted Unet3+, and the adapted Unet3+ includes 3 times of encoding and 3 times of decoding. During the test, the original image is input to the h1 node, generate a Curve Parameter Map, input it as a parameter into LE-curve to generate 4 enhanced images, and output the result of the hd1 branch.

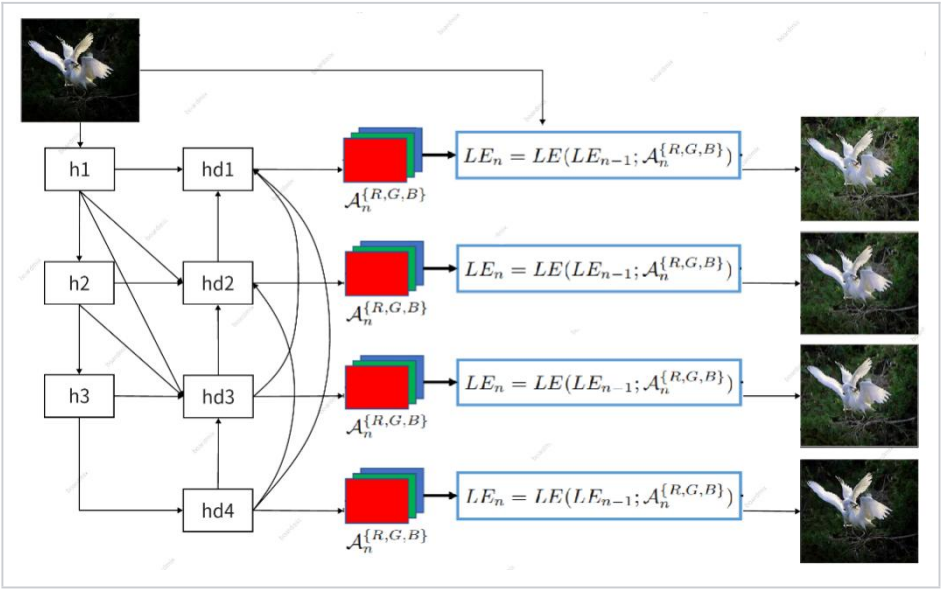


Fig. 2. Overall structure diagram of proposed method

2.1 Image pre-processing

When selecting the data set, the train set uses 3021 images of different exposure levels of SICE[7] Part1 and 360 images of ground truth exposure as the paired training set, which is cropped to 512 x 512 for easy access. The test set uses 767 weakly exposed images of SICE Part2 and their corresponding 229 ground truth exposure images as test data for quantitative metrics(PSNR/SSIM/MAE) ; 64 DICM and 10 LIME unlabeled data provided by Zero-DCE are used as test qualitative metrics (PIRM2018[8]) data.

2.2 Revisiting Unet3+

Unet3+ is a network model that was released in 2020, which is used in the network design to replace the Unet network in Zero-DCE. Deep supervision is one of its

characteristics. Unet3+ has four outputs when decoding. Each output is calculated and accumulated by weight when calculating the loss function. Unet3+ also has a large number of shortcut layers, which allows the network to fully consider the output of previous steps each time it decodes.

When designing Unet3+, to reduce computational complexity, the network was simplified to use only three encodings and three decodings, resulting in an adapted Unet3+ with four outputs. To perform fast computations, the network uses only convolutional and Relu layers. To ensure that the four outputs of Unet3+ have the same shape, the output channels of all convolutional layers are set to 24. The number 24 represents three output channels(RGB), each of which can go through 8 iterations of LE-Curve.

2.3 Squeeze and excitation

This paper adopts the SEnet [9] attention mechanism in Unet3+. Figure 3 depicts the overall method. It is a feature extraction model of deep learning, and its primary goal is to enhance the model's capacity to extract features. The central idea is to introduce a module called Squeeze-and-Exclusion, which learns the relationship between channels and adaptively adjusts the channel weights of feature graphs.

Using a global average pooling operation in the Squeeze stage, SENet reduces the input feature map from a three-dimensional tensor (height, width, and channel) to a one-dimensional vector. The per-channel average is calculated by the global average pooling operation, enabling the acquisition of a channel-specific feature description for each channel. SENet introduces a small multi-layer perceptron (MLP) in the Excitation stage, which has two completely connected layers. The goal of this MLP is to learn the relationship between channels by modeling the output of the Squeeze stage. The first fully connected layer, in particular, reduces the dimension of the input feature map before performing nonlinear transformation through an activation function (such as ReLU). The second completely connected layer then upscales the feature map and restores it to its original number of channels. The output is then restricted to the range of 0 to 1 by a Sigmoid function.

Through multiplication, the weight of each channel output in the Excitation stage is re-weighted to the original feature map. This implies that each channel's significance will be altered in accordance with its weight. SENet can now pay more attention to the feature channels that are more important for the current dark image enhancement task, such as the contour features of the input image, after re-weighting. The performance and generalizability of the model can be enhanced by this operation, which can make the entire network pay attention to the crucial features while suppressing the irrelevant ones. Network inference generates a Curve Parameter Map. This paper refers to Zero-DCE's LE-Curve stage, uses the Curve Parameter Map to process the original picture, and generates an enhanced picture.

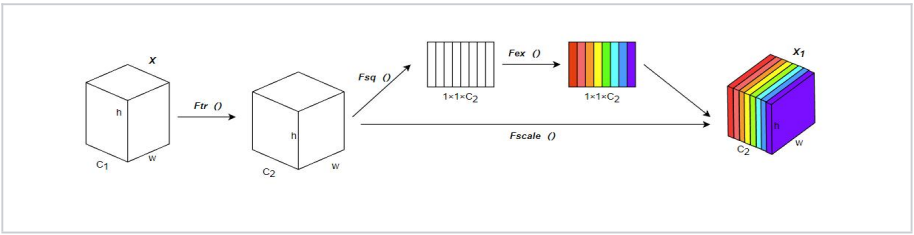


Fig. 3. SENet Attention Mechanism[9]

2.4 Image denoising

After forming the network, it was found that Unet3+ improved both the image and the noise, so the denoising module was used in this paper. The TV-Chambolle denoising method is used to denoise the image in this paper. The original image is loaded first in the processing process, and then Gaussian noise is added. This paper establishes appropriate parameters by estimating the standard deviation of noise. The noisy image is then denoise using the TV-Chambolle denoising function, and the result is converted to an 8-bit unsigned integer type. Finally, save the denoised image to the output folder you specified. This method can reduce noise while maintaining image details and improving image quality.

2.5 SSIM loss

The structural similarity index can evaluate the degree of distortion in images as well as the similarity of two images. SSIM is a perceptual model that, unlike MSE and PSNR, measures absolute error. It is more consistent with how human eyes naturally perceive things. The loss function focuses on three main features of an image: brightness, contrast, and structure. By adjusting the three parameters, and, the influence factor weights (α, β, γ) of the three influencing factors can be adjusted, and the enhanced image can be compared with the original label for learning, so that the resultant image's quality might be enhanced.

Based on the above analysis, considering the weakness of Zero-DCE in supervisory indexes PSNR, SSIM, and MAE during training, this paper introduces a supervisory SSIM loss, whose calculation method is shown in Equations (1) and (2). The brightness indices for the inputs X and Y are calculated as average brightness, contrast, and structure, respectively, and then compared to obtain the initial evaluation of similarity. To obtain the second evaluation, the contrast is calculated and compared after the impact of brightness has been subtracted. The control group has also been removed using the outcomes of the previous step, and the structure has been compared. The final evaluation result is created by combining the results.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x{}^2 + \mu_y{}^2 + C_1)(\sigma_x{}^2 + \sigma_y{}^2 + C_2)} \tag{1}$$

$$\mathrm{SSIMLoss} = 1 - \mathrm{SSIM}(x,y) \tag{2}$$

The total loss of our method is a linear combination of TV_loss, spa_loss, col_loss, exp_loss and ssim_loss, which can be seen in Equations (3).

$$\text{Loss} = 200 \times \text{TV\_loss} + 10 \times \text{spa\_loss} + 5 \times \text{col\_loss} + 10 \times \text{exp\_loss} + 100 \times \text{ssim\_loss} \quad (3)$$

2.6 Model training

Pairing data, which comprises of the original picture and the improved ground truth picture, should be fed into the Unet3+ network during training to produce four Curve Parameter Maps. The LE function calculates the four maps to generate four enhanced images. The Zero-DCE loss functions (space consistency error, exposure control error, color constant error, illumination smoothing error) are calculated directly from the enhanced pictures, while the SSIM loss is calculated by comparing the enhanced ground truth image to the enhanced images. These loss functions are accumulated based on their weight, and network training is performed.

3 Experiments

3.1 Datasets

This study employs a training set consisting of 3021 images from part 1 of the SICE dataset, accompanied by their corresponding label images. The multi-exposure enhanced photos in the SICE dataset were taken in a variety of settings and at various times. To facilitate data processing, the images were resized to 512x512 dimensions and stored. To assess the performance of the network model, two test sets were devised in line with the methodology outlined in Zero-DCE paper. The supervised indicators of the network model were evaluated using part 2 of the SICE dataset, whereas the unsupervised indicators were assessed based on 10 LIME and 64 DICM images.

3.2 Evaluation metrics

In this study, the unsupervised learning method proposed in the PIRM paper is adopted, which integrates two unsupervised indicators, Ma [10] and NIQE [11]. NIQE is an unsupervised learning index for evaluating the quality of images. It utilizes a set of local statistical features, including the image's gradient, contrast, brightness, and texture information. Compared to traditional evaluation metrics like PSNR or SSIM, NIQE places greater emphasis on the perceptual quality of the image, rather than solely focusing on pixel-level differences. It provides a better reflection of how the human eye perceives image quality. Ma is a non-reference metric that was learned from scores on visual perception. The index includes three low-level features in the space and frequency domain that are used to quantify super-resolution artifacts. Then, without using real ground photos, a two-stage regression model is developed to forecast the quality score of high-resolution photographs.

The final unsupervised index PIRM used in this study is (NIQE+(10-Ma))/2, and the lower the value, the higher the image quality. This is done to strike a balance between the trade-off between image quality and model performance and help the model achieve a better comprehensive evaluation in the dark image enhancement task.

Three related metrics, PSNR, SSIM, and MAE, are used in supervised learning. Peak Signal-to-Noise Ratio (PSNR), a commonly used statistic for evaluating image quality, is used to assess how much noise and signal there is in an image and how well a denoising method is working. The image quality increases as the value increases. The Structural Similarity Index (SSIM) is an index used to assess the quality of photographs and determine how similar two images are to one another. This parameter index assumes that the structural information in the image should remain unchanged when it is not distorted and takes into account the similarity of brightness, contrast, and structure as well as the subjective perception of human eyes. Because of this, SSIM compares the structural information of the two images to determine how similar they are. The mean absolute error (MAE) index calculates the average discrepancy between the expected value and the actual value. This loss is more resilient and unaffected by extreme values as compared to other often used indicators, such as mean square error (MSE). Additionally, because MAE only considers the absolute value of the error, it can more accurately reflect the error in real-world situations.

3.3 Experiment settings

All testing was carried out on a machine with an Intel Core i7 CPU and 16GB of RAM. The PyTorch deep learning framework is used in this study to create the model network, which is trained and tested on an NVIDIA GeForce GTX 2080Ti GPU. Additionally, the learning rate is set at 0.0001 and its gradient decrements by 10 times per 20 epochs.

The paired training set is first input into the 256x256 network model, which generates an enhanced picture. The color loss, exposure loss, illumination loss, and space loss are then calculated for the enhanced picture itself. The SSIM loss is calculated using the enhanced pictures and label pictures, and the total loss is obtained by adding them in a specified proportion. After that, the model is trained using the Adam optimizer.

3.4 Performance analysis

The metric this paper used are reported in Table 1, a lower PIRM value means better perceptual quality, the same as MAE. In terms of qualitative indicators, we performed better than Zero-DCE on the DICM dataset. And, in terms of quantitative indicators, all of our indicators(PSNR/SSIM/MAE) are better than Zero-DCE.

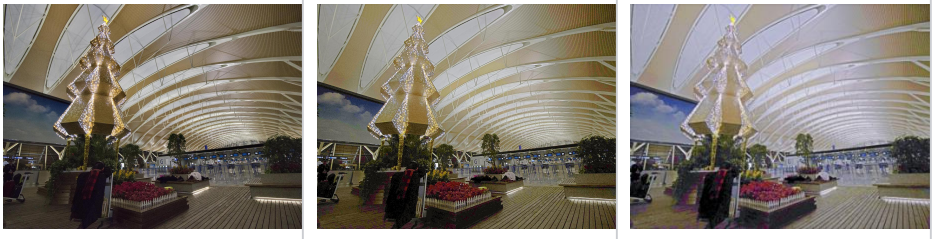
Table 1. Model performance comparison between Zero-DCE and this paper

	LIME PIRM metric	DICM PIRM metric	SICE PSNR metric	SICE SSIM metric	SICE MAE metric
Zero-DCE	2.76	3.04	16.57	0.59	98.78
ours	2.8785	3.0277	18.1590	0.5998	81.0976

Numerous ablation tests are conducted in this research to demonstrate the effectiveness of each component. As a supervised loss function, SSIMLoss aims to make the output images of the network model more closely resemble the label images in terms of brightness, contrast, and structure. The comparison results between the model with SSIMLoss and the original model are presented in the Table 2 and Figure 4 below. The inclusion of SSIMLoss resulted in a substantial drop in the unsupervised index and a rise in the supervised index, indicating that the enhanced images achieved better brightness and contrast while maintaining image quality.

Table 2. Model performance with/without SSIMLoss

	LIME PIRM metric	DICM PIRM metric	SICE PSNR metric	SICE SSIM metric	SICE MAE metric
original result	2.8785	3.0277	18.1590	0.5998	81.0976
w/o SSIM Loss	3.2423	3.3333	16.5723	0.5783	101.8291



(a) Ground truth[7] (b) Our enhancement (c) Without SSIM Loss

Fig. 4. Visual comparison of enhancement with/without SSIM Loss

To further examine the role of SEBlock, we conduct a number of tests, and the outcomes are displayed in Table 3 and Figure 5. As an attention mechanism, SEBlock increases the network's width and adaptability. The comparison between the results of removing SEBlock and the original model is shown in the table below. While SEBlock's unsupervised index has largely remained unchanged, its supervised learning index has clearly improved. This demonstrates that while the SEBlock network's ability to fit data has undoubtedly improved, the quality of the image has not. To improve the model's ability to perceive dark details, add SEBlock attention

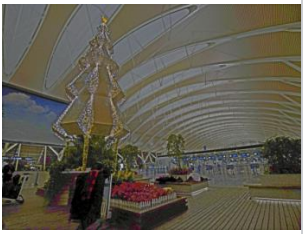
mechanisms at the entrance of each input channel. The model will be better able to enhance subtle texture, edge, and detail information as a result of this, and it will also be better able to change the image's brightness distribution, resulting in a more balanced brightness change.

Table 3. Model performance with/without SENet

	LIME PIRM metric	DICM PIRM metric	SICE PSNR metric	SICE SSIM metric	SICE MAE metric
this paper	2.8785	3.0277	18.1590	0.5998	81.0976
w/o SENet	2.8128	3.0758	15.4659	0.5641	106.7277

(a) Ground truth[7]

(b) Our enhancement



(c) Without SENet

Fig. 5. Visual comparison of enhancement with/without SENet

In order to address the presence of noticeable noise in some of the images when examining the network results, this paper attempts to utilize a simple denoising module called TV-Chambolle to mitigate the noise issue. By reducing the image's overall fluctuation, the TV-Chambolle method lowers the amount of noise in the image.To be more specific, after loading the image to be processed, a random noise function is used to add Gaussian noise to the original picture to simulate the impact of noise in the actual world. The estimation function is then used to estimate the noise standard deviation in order to determine the noise intensity. The image with noise added is then denoised using the TV-Chambolle algorithm. The algorithm gradually reduces the total variation of the image through iterative optimization in order to produce the desired denoising effect. The updated image's pixel value is determined according to the gradient in each iteration. The denoised image is finally obtained.

As Table 4 and Figure 6 shown, the TV-Chambolle algorithm-based method of image denoising can reduce noise to boost the clarity of low-light photographs. It should be noted that while using this method, and the supervised indicators may be improved, the unsupervised indicators may be decreased. This is due to the fact that part of the image's fine details may be lost during the denoising process, which can also have an impact on the performance of some unsupervised indicators.

Table 4. Model performance with/without denoising

	LIME PIRM metric	DICM PIRM metric	SICE PSNR metric	SICE SSIM metric	SICE MAE metric
this paper	2.8785	3.0277	18.1590	0.5998	81.0976
with denoising	3.1207	3.1260	15.2183	0.6018	80.2510

(a) Ground truth[7] (b) Our enhancement (c) with denoising

Fig. 6. Visual comparison of enhancement with/without denoising

4 Discussion

In this study, a technique for improving low-light photographs using the Unet3+ network is proposed and the SEnet attention mechanism, training with the SSIM loss function, and introducing a denoising module for improved results. There are still areas for optimization and enhancement in this study, just like there are in other techniques for boosting low-light photographs. These improvements can be explored in areas such as data augmentation, model architecture, loss functions, training strategies, and denoising techniques.

(1) In terms of data augmentation, the current paper only focuses on enhancing low-light images and does not consider optimizing overexposed images. However, in practical applications, overexposed images are frequently encountered. Therefore, a promising approach for improvement is training a model that optimizes both overexposed and underexposed images simultaneously. By combining these two scenarios, the optimized images can have more stable brightness, which improves the robustness of the enhancement results.

(2) In terms of model architecture, the paper chooses the Unet3+ network as the basic framework, supplemented by the SEnet attention mechanism. Despite the good performance achieved in this paper, overfitting is still a common problem in practical applications. In order to overcome overfitting, additional regularization mechanisms can be considered, such as adding regularization items or using Dropout[12] layers to reduce model complexity and improve generalization ability. Furthermore, different model architectures can be explored, such as introducing residual connections or adopting deeper network structures, to further enhance the enhancement effect.

(3) SSIMLoss is chosen as the loss function for supervised learning in terms of loss function selection. Although SSIMLoss has benefits when taking structural similarity and perceived quality into account together, it is not always the best option. We can experiment with designing different loss functions to further enhance the enhancement effect. For instance, using a pre-trained feature extraction network and perceptual loss, it is possible to compare the perceptual differences between the generated image and the target image. It is also possible to take into account KL divergence loss, which measures the disparity between the distributions of the generated and target images and can be used to make the generated image distributions more similar to the label distributions of the target image.

(4) In terms of training strategies, the paper employs more than 100 pairs of different loss weight combinations as an effective strategy for exploring loss function weight combinations. However, this method does not combine the weights for multi-supervision in Unet3+. Further improvements can be attempted by trying different weight combination strategies, including different combinations for the attention mechanism in Unet3+. The model's ability to enhance itself and its speed of convergence can both be further enhanced by optimizing the training strategy.

(5) In terms of denoising, the paper introduces a denoising module to improve low-light image enhancement. However, this module has some limitations. The current implementation uses a widely used traditional denoising algorithm in the field of image processing. The fundamental principle of this algorithm involves smoothing the image to reduce noise. However, the algorithm employs fixed parameter values (sigma and weight), which may result in inconsistent denoising outcomes across different images. Since different images may have different noise levels and noise characteristics, more flexible denoising methods are required to adapt to different scenes and noise types. In future research, more advanced denoising algorithms, such as deep learning-based methods or adaptive denoising techniques, can be explored to enhance the robustness and adaptability of the denoising effect. These enhancements would considerably enhance the performance and utility of the proposed technique for improving low-light photographs.

5 Conclusion

In this paper, we propose a low-light image enhancement (LLIE) method based on image denoising and structural similarity loss, aiming at alleviating the noise from the black color blocks and halo near the boundary of the bright area. We first transform DCE-Net in Zero-DCE to Unet3+, which enhances the network's fitting ability. Then, we introduce a denoising module and an SSIM loss, which can improve the qualitative and quantitative metrics of the network. Extensive experimental findings show that our suggested approach works. We finally discuss the possible directions to improve the quality and universality of the enhancement effect from designing model architecture, selecting an appropriate loss function, optimizing the training strategy, and enhancing the denoising module.

Acknowledgment

Qiyao Li is responsible for Abstract, Keywords, Literature review and Conclusion. Zhequan Li is responsible for Introduction, Results, Discussion and References. Haoyang Wang is responsible for Method and the conduct of the experiment.

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Lore, K. G., Akintayo, A., & Sarkar, S.: LLNet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition, 61, 650-662 (2017).
2. Wei, C., Wang, W., Yang, W., & Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560. (2018)
3. Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., & Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1780-1789) (2020).
4. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., ... & Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE transactions on image processing, 30, 2340-2349 (2021).
5. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., ... & Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1055-1059). IEEE (2020, May).
6. Zhao, H., Gallo, O., Frosio, I., & Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on computational imaging, 3(1), 47-57 (2016).
7. Cai, J., Gu, S., & Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing, 27(4), 2049-2062 (2018).
8. Blau, Y., & Michaeli, T.: The perception-distortion tradeoff. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6228-6237) (2018).
9. Hu, J., Shen, L., & Sun, G.: Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141) (2018).
10. Ma, C., Yang, C. Y., Yang, X., & Yang, M. H.: Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding, 158, 1-16 (2017).
11. Mittal, A., Soundararajan, R., & Bovik, A. C.: Making a “completely blind” image quality analyzer. IEEE Signal processing letters, 20(3), 209-212 (2012).
12. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958 (2014).