Low-Light Image Enhancement Network Based on Multi-Scale Feature Complementation

Yong Yang[1*], Wenzhi Xu[2*], Shuying Huang[3†], Weiguo Wan[4]

[1]School of Computer Science and Technology, Tiangong University, Tianjin, China

[2]School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China

[3]School of Software, Tiangong University, Tianjin, China

[4]School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang, China[4]

greatyangy@126.com, xuwenzhi1998@163.com, huangshuying@tiangong.edu.cn, wanweiguo@jxufe.edu.cn

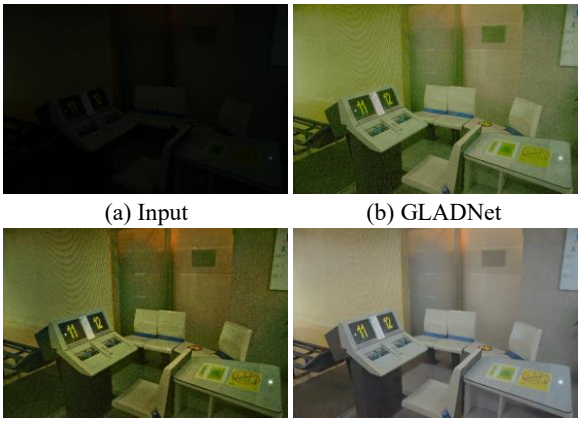
Abstract

Images captured in low-light environments have problems of insufficient brightness and low contrast, which will affect subsequent image processing tasks. Although most current enhancement methods can obtain high-contrast images, they still suffer from noise amplification and color distortion. To address these issues, this paper proposes a low-light image enhancement network based on multi-scale feature complementation (LIEN-MFC), which is a U-shaped encoder-decoder network supervised by multiple images of different scales. In the encoder, four feature extraction branches are constructed to extract features of low-light images at different scales. In the decoder, to ensure the integrity of the learned features at each scale, a feature supplementary fusion module (FSFM) is proposed to complement and integrate features from different branches of the encoder and decoder. In addition, a feature restoration module (FRM) and an image reconstruction module (IRM) are built in each branch to reconstruct the restored features and output enhanced images. To better train the network, a joint loss function is defined, in which a discriminative loss term is designed to ensure that the enhanced results better meet the visual properties of the human eye. Extensive experiments on benchmark datasets show that the proposed method outperforms some state-of-the-art methods subjectively and objectively.

Introduction

Images captured in extremely dark environments often have problems such as low brightness, low contrast, severe noise pollution, and color loss, which will affect subsequent image processing tasks, such as security monitoring, target detection and recognition. Therefore, effective low-light image enhancement methods are needed to restore the quality of low-light images.

In recent years, low-light image enhancement methods based on deep learning have gradually become mainstream due to the powerful feature extraction capabilities of deep convolution networks. These methods are roughly divided into three categories according to the learning strategy of the network: supervised learning, semi-supervised learning and unsupervised learning. The earliest methods are mainly based on supervised learning networks. For example, LLNet (Lore, Akintayo and Sarkar 2017) is a supervised learning network and the first neural network built for low-light image enhancement, in which a stacked sparse auto-encoder is proposed for noise removal and contrast enhancement. To reduce the network's dependence on training samples, semi-supervised and unsupervised learning networks are proposed. DRBN (Yang et al. 2021) is a deep recursive band network that combines fully supervised and unsupervised learning. It first used supervised learning to recover linear bands of the enhanced image and then recombined these bands through unsupervised adversarial learning to obtain an improved band representation. Unsupervised learning does not require paired datasets to participate in network training. For example, EnlightenGAN (EG) (Jiang et al. 2021) is the first unsupervised low-light



(a) Input (b) GLADNet (c) EnlightenGAN (d) Ours

Figure 1: Results of different low-light image enhancement methods. GLADNet (Wang et al. 2018) and EnlightenGAN (Jiang et al. 2021) tend to produce images with strong noise.

*These authors contributed equally. †Corresponding author.

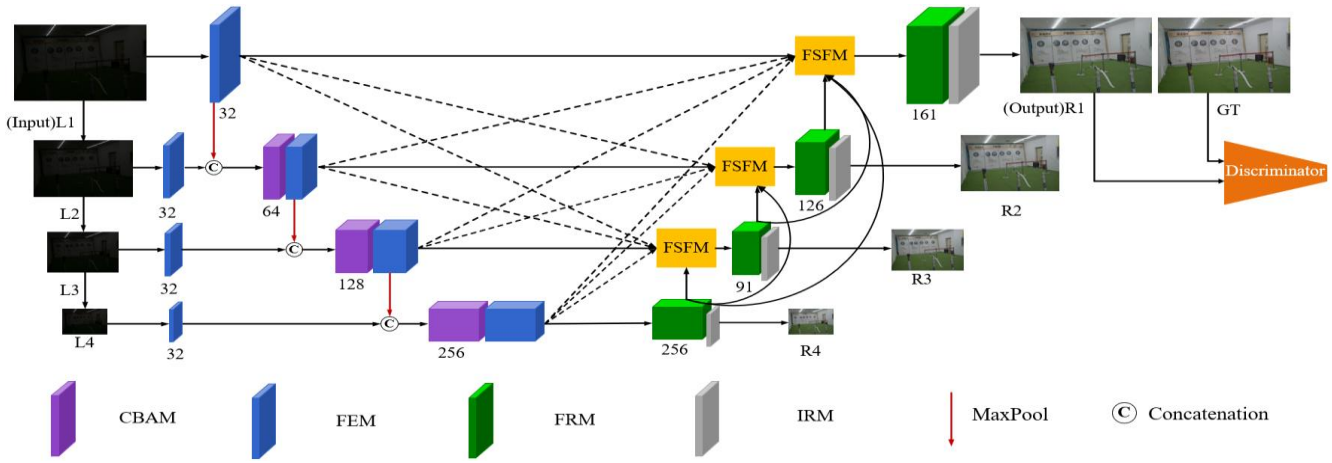


Figure 2: The overall framework of LIEN-MFC.

image enhancement network, which uses a U-Net network as a generator and a global-local discriminator to enhance low-light images. RetinexDIP (Zhao et al. 2022) combined Retinex theory with deep learning and built two encoders to generate a reflection map and an illumination map from randomly sampled white noise, and used the illumination map to solve the final enhanced result according to Retinex theory. Although deep learning-based methods have made great progress, they are not suitable for all images obtained in low-light scenes, and the obtained enhancement results still suffer from noise amplification and color distortion, as shown in Figure 1.

In response to the current problems, we propose a low-light image enhancement network based on multi-scale feature complementation (LIEN-MFC), which is an encoder-decoder network with a U-shaped structure. The encoder performs feature extraction for low-light images at multiple scales. The decoder complements and fuses the multi-scale features extracted by the encoder and the decoder to realize the reconstruction of features at each scale. To better restore the color and content of the image, a joint loss function is defined to train the network. The contributions of this paper are as following.

In LIEN-MFC, an encoder-decoder structure with four branches, is proposed to enhance low-light images by making full use of the multi-scale features extracted by the network.

A feature supplementary fusion module (FSFM) is constructed and used in the encoder to integrate the features from multiple scales of the encoder and decoder.

A joint loss function consisting of multiple loss terms is defined, in which an adversarial loss term and a saturation loss term are designed to make the enhanced image more visually natural and reduce the degree of color distortion.

Proposed Approach

In this section, a LIEN-MFC is proposed to realize the low-light image enhancement, as shown in Figure 2, which

Figure 3: Specific structure of some modules in Figure 2.

is designed as an encoder-decoder network with a U-shaped structure. Figure 3 shows the specific structure of some modules in LIEN-MFC. The encoder is constructed to extract features from low-light images at multiple scales, while the decoder aims to reconstruct the features of the image by implementing feature complementarity at different scales. To make the recovered images more in line with the visual properties of the human eye, a joint loss function is defined to guide the network training. Next, we describe the structure of LIEN-MFC and the definition of the loss function in detail.

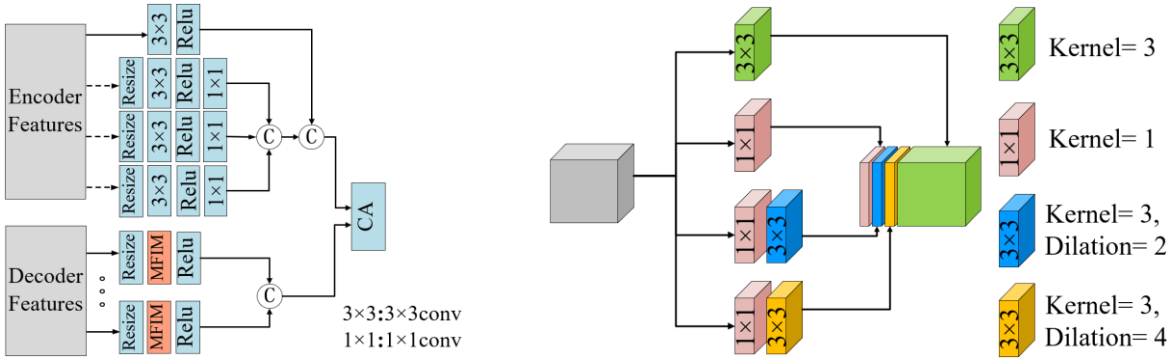
Structure of LIEN-MFC

The proposed LIEN-MFC is designed as a multi-supervised encoder-decoder network with four branches, each of which is used to extract and restore features at one scale. Therefore, the LIEN-MFC takes four low-light images of different scales as input, which are obtained by down-sampling the original-sized low-light image. Likewise, to ensure the accuracy of restored features at each scale, the corresponding ground-truth (GT) images are also downsampled to supervise the output of each branch. The low-light images of different scales can be obtained by the following equation.

$$L_{n+1} = Down(L_1, 2^n) \quad (1)$$

where L_1 is the input image, and L_n (n=1,2,3,4) denotes the low-light images at different scales. $Down(\cdot)$ represents the downsampling operation, and 2^n (n=1,2,3) represents the downsampling factor.

In the encoder of LIEN-MFC, two feature extraction modules are constructed and used, namely, feature extrac-



(a) (b)

Figure 4: The components of FSFM. (a) The structure of FSFM. (b) The structure of MFIM in FSFM.

tion module (FEM) and convolutional block attention module (CBAM) (Woo et al. 2018). In each branch, a FEM is utilized to extract shallow features from low-light images at corresponding scales. In the second, third and fourth branches, a CBAM and a FEM are further used for deeper feature extraction and enhancement. To increase the completeness of the features extracted from each branch, the shallow features of each branch are integrated into the next scale branch through a maxpooling operation as supplementary information. The processing of encoder can be expressed by the following equations:

$$\begin{aligned} EnO_1 &= FE(L_1) \\ EnO_n &= FE(CBA(Cat(MP(EnO_{n-1}), FE(L_n)))) \end{aligned} \tag{2}$$

where EnO_n (n=1,2,3,4) represents the output feature maps of each branch in the encoder, $FE(\cdot)$ represents the operation of FEM consisting of two convolutional layers, and $CBA(\cdot)$ represents the operation of CBAM. In addition, $Cat(\cdot)$ and $MP(\cdot)$ represent the concatenation operation and the maximum pooling operation, respectively.

In the decoder, all four branches contain a feature recovery module (FRM) and an image reconstruction module (IRM) to achieve feature recovery and image reconstruction at each scale. The FRM consists of five residual blocks (ResBs) connected in series, and the IRM contains two convolutions and a ReLU operation to realize the mapping of feature maps to enhanced images. To better utilize the features of each scale, a FSFM is constructed and used in the first, second and third branches to integrate features from different scales in the decoder and the reconstructed features from the encoder. The process of decoder can be represented by the following equations.

$$DeO_4 = FRM(EnO_4), R_4 = IRM(DeO_4) \tag{3}$$

$$\begin{aligned} DeO_3 = FRM(&FSFM(EnO_1, EnO_2, EnO_3, EnO_4, \\ &Up(DeO_4))), \end{aligned} \tag{4}$$

$$R_3 = IRM(DeO_3)$$

$$\begin{aligned} DeO_2 = FRM(&FSFM(EnO_1, EnO_2, EnO_3, EnO_4, \\ &Up(DeO_3), Up(DeO_4))), \end{aligned} \tag{5}$$

$$R_2 = IRM(DeO_2)$$

$$\begin{aligned} DeO_1 = FRM(&FSFM(EnO_1, EnO_2, EnO_3, EnO_4, \\ &Up(DeO_2), Up(DeO_3), Up(DeO_4))), \\ R_1 = IRM(&DeO_1) \end{aligned} \tag{6}$$

where DeO_n represents the output of FRM in each branch in the decoder, $FSFM(\cdot)$ represents the operation of FSFM, $FRM(\cdot)$ represents the operation of FRM, R_n (n=1,2,3,4) represents the reconstructed image of each branch, and R_1 is the final enhanced image. Below we describe the structure of FSFM in detail.

Feature Supplementation Fusion Module (FSFM)

Considering the difference between the features from the encoder and decoder, a FSFM is constructed, as shown in Figure 4 (a). It first integrates features of different scales from encoder and decoder, respectively, and then further fuses and enhances the integrated features of the two parts by a channel attention (CA) module (Hu et al. 2020). The feature integration process of FSFM is described as follows.

For the features from the encoder, since they come from different branches and have different sizes, they need to be resized to the same size as the feature maps under the current branch. Then, a convolution layer and a 1×1 convolution operation are used to further extract features and reduce the dimension of the feature maps in the channel direction. For the feature maps from the current branch, their size and number of channels remains unchanged, and only one convolutional layer is used for feature extraction. Finally, two concatenation operations are employed to achieve the integration of the features from different branches.

For the features of the decoder, they come from different branches of FRMs. And these features, especially those from the third and second branches, have rich shallow and deep semantic features. Therefore, to learn rich features, a multi-scale feature integration module (MFIM) is designed to extract and integrate features of different scales using convolution kernels with different receptive fields. The structure of MFIM is shown in Figure 4 (b), which uses convolution kernels of different sizes and dilation rates to

extract features of different scales. The extracted feature maps are dimensionally reduced by 1×1 convolution to generate a feature map to complement the features at the current scale. Here, the branch with only one 3×3 convolution keeps the channels of the feature map unchanged to preserve the main information of the current scale. Finally, these features are combined through a concatenation operation and fed into the CA module together with the integrated features from the encoder to obtain the output of FSFM.

Loss Function

To better constrain the enhanced images to be closer to the GT images, we define a joint loss function, which consists of five loss terms: content loss, structural loss, perceptual loss, saturation loss, and adversarial loss. It can be defined as follows.

$$L_{total} = \alpha_1 L_c + \alpha_2 L_s + \alpha_3 L_p + \alpha_4 L_{sa} + \alpha_5 L_a \tag{7}$$

where L_c, L_s, L_p, L_{sa}, and L_a represent the content loss term, structural loss term, perceptual loss term, saturation loss term, and adversarial loss term, respectively. In addition, α_1, α_2, α_3, α_4, and α_5 are the weights of different loss terms.

Content Loss. The content loss consists of a smooth $L1$ loss and a cosine similarity loss, which mainly ensures that the content of the recovered image at each scale is closer to that of the GT image at the corresponding scale. The content loss term is defined as follows.

$$L_c = \sum_{i=1}^{4} \lambda_i \left(SmoothL1(R_i, GT_i) - Cosine(R_i, GT_i) \right) \tag{8}$$

where λ_i denotes the weight of content loss term at the i-th scale, R_i denotes the enhancement result at the i-th scale, and GT_i denotes the GT image at the i-th scale. $SmoothL1(\cdot)$ denotes the smooth $L1$ loss, and $Cosine(\cdot)$ denotes the cosine similarity loss.

Structural Loss. Structural similarity (SSIM) loss (Lv, Liu and Lu 2020) is defined to measure the structural similarity between images in terms of luminance, contrast, and structure, and is defined as follows.

$$L_s = 1 - SSIM\left(R_1, GT_1\right) \tag{9}$$

where $SSIM(\cdot)$ represents the function that calculates the SSIM metric of two images.

Perceptual Loss. Perceptual loss (Xu et al. 2020) is a method to measure image similarity by comparing the difference of features between two images, which more accurately simulates human perception of images. Therefore, this work constructs a perceptual loss term using the feature maps extracted by the first and second layers of the pretrained VGG-16, which is defined as follows.

$$L_p = \left\| \phi\left(R_1\right) - \phi\left(GT_1\right) \right\|_2^2 \tag{10}$$

where $\phi(\cdot)$ represents the feature maps from the first and second convolution layers of VGG-16.

Saturation Loss. To ensure that the color of the recovered images is closer to that of the GT images, we propose a saturation loss term, which is defined by computing the color moments of the image that describe the color distribution. Since the color distortion is mainly caused by the inaccurate restoration of color saturation, this paper defines the saturation loss term in the HSV color space, which is expressed as follows.

$$L_{sa} = \left| \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} S_{i,j}^{R} - \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} S_{i,j}^{T} \right| \tag{11}$$

where i, j denote the indexes of pixel positions. H, W are the height and width of images. S^T is the saturation component of GT_1 and S^R is the saturation component of R_1.

Adversarial Loss. To make the distribution of the enhanced image closer to that of the GT image, we define an adversarial loss term (Afifi et al. 2021) to ensure that the enhanced image is more visually realistic. Here, the discriminator in PatchGAN (Demir and Unal 2018) is used and trained together with the proposed network to constrain the enhanced results of the network. The adversarial loss term is defined as follows.

$$L_a = -3HWNlog(Sigmod(PD(R_1))) \tag{12}$$

where $Sigmod(\cdot)$ is the Sigmoid function and $PD(\cdot)$ is the discriminator of PatchGAN. N represents the number of branches in LIEN-MFC.

Experimental Results and Analysis

Implementation Details

During training, a two-stage training strategy is adopted to train the proposed network. The first stage sets the number of iterations to 250 and uses only the content loss, perceptual loss, structure loss and saturation loss to train the network. After obtaining better enhanced results in the first stage, the second stage adds the adversarial loss to further optimize the network. Through extensive experiments, the weights α_1, α_2, α_3, α_4, and α_5 in Eq. (8) were set to 1, 0.25, 0.25, 0.1, and 1, respectively. The weights λ_1, λ_2, λ_3, and λ_4 in Eq. (9) were set to 2, 1, 1, and 1, respectively. In addition, the initial learning rate is set to 0.0002, and when the number of iterations reaches 150, the learning rate is adjusted to half of the initial learning rate.

Dataset and Metrics

In experiments, the LOL (Wei et al. 2018) dataset is used to train the proposed network, which consists of 500 low-light and normal-light image pairs, 485 of which are used for training and 15 for testing. To test the generalization ability of the network, the trained network is also tested on another dataset VE-LOL-L (Liu et al. 2021).

To quantitatively evaluate the performance of various networks, the peak signal to noise ratio (PSNR), SSIM

(a) Input (b) GLADNet (c) EG (d) KIND (e) ZERO-DCE

(f) DLN (g) KIND++ (h) URetinex (i) Ours (j) GT

Figure 5: Subjective comparison with state-of-the-art low-light image enhancement methods on LOL dataset.

(a) Input (b) GLADNet (c) EG (d) KIND (e) ZERO-DCE

(f) DLN (g) KIND++ (h) URetinex (i) Ours (j) GT

Figure 6: Subjective comparison with state-of-the-art low-light image enhancement methods on VE-LOL-L dataset.

(Wang et al. 2004), and learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018) are used as the evaluation metrics. The larger the values of PSNR and SSIM, the more similar the images are. LPIPS measures the difference between two images by learning the similarity of perceived image blocks, which is more in line with human perception than traditional indicators. The smaller the value of LPIPS, the more similar the images are.

Objective and Subjective Comparison

To demonstrate the effectiveness of the proposed method, we objectively compare the PSNR, SSIM and LPIPS values of the results obtained by the proposed method with those obtained by 13 state-of-the-art methods, including NPE (Wang et al. 2013), BIMEF (Ying, Li and Gao 2017), LIME (Guo, Li and Ling 2017), Dong (Dong et al. 2011), MF (Fu et al. 2016a), SRIE (Fu et al. 2016b), GLADNet (Wang et al. 2018), EG, KIND (Zhang, Zhang and Guo 2019), etc. The results obtained by the comparison methods are from the official codes.

Table 1 and Table 2 show the objective average metrics of all the compared methods on the LOL and VE-LOL-L datasets, respectively. As can be seen from Table 1 and Table 2, all quantitative metrics obtained by the proposed method are significantly better than those obtained by other state-of-the-art methods. This indicates that the proposed method can recover images that are closer to GT images.

Figure 5 shows the enhancement results for one low-light image in the LOL dataset, and Figure 6 shows the enhancement results for one low-light image in the VE-LOL-L dataset. Only the results of the deep learning comparison methods are presented here. As can be seen from the figures, the results obtained by GLADNet have severe color casts and noise. The results of EG and Zero-DCE also have some noise, and the results of KIND have blurred edges. And the above four methods all have the problem of underexposure. The results obtained by DLN, KIND++, and URetinex suffer from color distortions and artifacts compared to the GT images. The results of the proposed method are closer to the GT images in both color and texture structure.

Ablation Study

Effect of Loss Function. To investigate the effect of different loss terms on network training, we designed five sets of experiments to test the effect of different loss terms on

Methods	PSNR↑	SSIM↑	LPIPS↓
NPE (Wang et al. 2013)	16.970	0.579	0.405
BIMEF (Ying, Li and Gao 2017)	13.875	0.679	0.326
LIME (Guo, Li and Ling 2017)	16.920	0.595	0.360
Dong (Dong et al. 2011)	16.717	0.572	0.385
MF (Fu et al. 2016a)	16.964	0.602	0.380
SRIE (Fu et al. 2016b)	11.855	0.573	0.340
GLADNet (Wang et al. 2018)	19.718	0.714	0.321
EG (Jiang et al. 2021)	17.484	0.699	0.322
KIND (Zhang, Zhang and Guo 2019)	20.379	0.878	0.159
Zero-DCE (Guo et al. 2020)	17.330	0.632	0.387
DLN (Wang et al. 2020)	<u>21.944</u>	<u>0.900</u>	0.142
KIND++ (Zhang et al. 2021)	21.803	0.878	0.158
URetinex (Wu et al. 2022)	21.328	0.880	<u>0.121</u>
Ours	**24.732**	**0.912**	**0.107**

Table 1: Objective average metrics on the LOL dataset. ↑ (↓) denotes that the larger (smaller) the value, the better the quality. The best and second-best results are highlighted in bold and underlined, respectively.

Methods	PSNR↑	SSIM↑	LPIPS↓
NPE (Wang et al. 2013)	17.332	0.553	0.421
BIMEF (Ying, Li and Gao 2017)	17.855	0.713	0.286
LIME (Guo, Li and Ling 2017)	15.241	0.503	0.415
Dong (Dong et al. 2011)	17.255	0.568	0.385
MF (Fu et al. 2016a)	18.726	0.604	0.388
SRIE (Fu et al. 2016b)	14.451	0.601	0.312
GLADNet (Wang et al. 2018)	17.667	0.702	0.373
EG (Jiang et al. 2021)	18.640	0.714	0.309
KIND (Zhang, Zhang and Guo 2019)	23.783	0.887	0.121
Zero-DCE (Guo et al. 2020)	18.059	0.665	0.313
DLN (Wang et al. 2020)	<u>25.610</u>	<u>0.903</u>	0.115
KIND++ (Zhang et al. 2021)	22.211	0.859	0.175
URetinex (Wu et al. 2022)	21.221	0.868	<u>0.099</u>
Ours	**29.647**	**0.919**	**0.070**

Table 2: Objective average metrics on the VE-LOL-L dataset.

network performance, and the experimental results are presented in Table 3. Table 3 shows the results obtained by the trained network when one loss term is removed from the joint loss function L_{total} . From the table, we can see that the performance of the network degrades when a certain loss term is removed. The L_c and L_s terms respectively ensures that the content and structure of the enhanced image are close to those of the GT image, so there is a significant drop in the metrics of the results obtained by the network after removing these terms. In addition, after removing the L_{sa} term, the performance of the network is also

	PSNR↑	SSIM↑
w/o L_s	23.829	0.892
w/o L_p	<u>24.155</u>	0.908
w/o L_{sa}	24.043	<u>0.909</u>
w/o L_a	24.480	0.907
w/o L_c	18.427	0.820
Ours	**24.732**	**0.912**

Table 3: Quantitative comparison of different loss functions on the LOL dataset.

Methods	PSNR↑	SSIM↑
w/o FSFM	19.057	0.872
w/o MFIM in FSFM	24.287	0.904
w/ FSFM, w/ MFIM in Decoder (Ours)	**24.732**	**0.912**

Table 4: Ablation Study of FSFM and MFIM.

significantly reduced, which indicates that the defined L_{sa} term is helpful for image restoration.

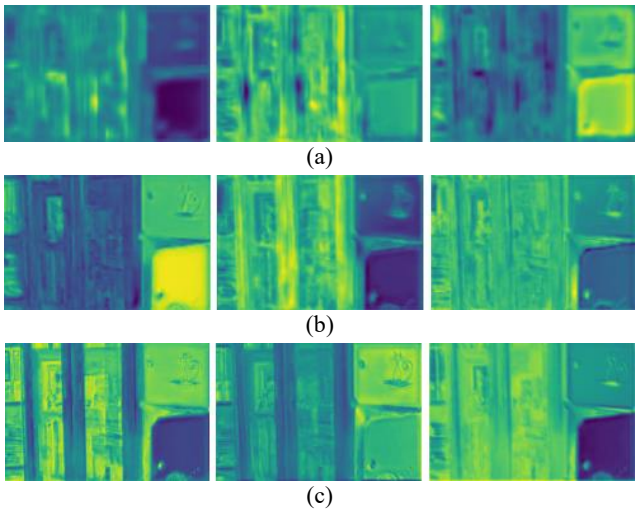
Effect of FSFM. FSFM is the core component of the proposed network, which fully fuses the multi-scale feature information generated by encoder and decoder. Here, we design ablation experiments from two perspectives to verify the effectiveness of FSFM.

One experiment is to evaluate the contribution of FSFM to the network by removing FSFM and replacing it with a concatenation operation. Another experiment is to evaluate the contribution of MFIM to the network by replacing MFIM in FSFM with a convolution layer. Table 4 shows the results obtained by the network without FSFM, network without MFIM in FSFM, and network with both. Obviously, after removing FSFM from the network, the performance of the network is greatly reduced. The performance of the network also degrades after removing MFIM in FSFM. Figure 7 shows the subjective results for several cases in Table 4. We can intuitively see that the model without FSFM or without MFIM in FSFM leads to severe color distortion and more artifacts in the enhancement results. The proposed model with FSFM and MFIM can obtain the enhanced result with more realistic colors,which is closer to the GT image.

When the MIFMs in FSFM are removed, although the performance of the model is substantially improved compared to the model without FSFM, there are still some gaps compared to the final model. To better demonstrate the role of MFIM in the network, three supplement feature maps from MFIM in the first branch are visualized as shown in Figure 8. The three feature maps in Figure 8 (a), (b), and (c) show the feature maps from a 1×1 convolutional layer, a convolutional layer with a dilation rate of 2 and a convolutional layer with a dilation rate of 4. It can be seen that convolutions with different receptive fields extract differ-

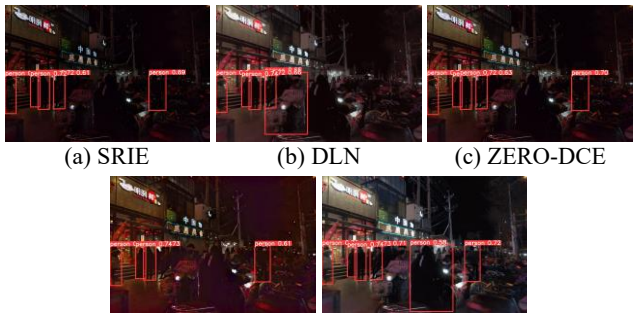
(a) (b) (c) (d) (e)

Figure 7: Illustration of the validity of FSFM and MFIM. (a) input. (b) w/o FSFM. (c) w/ FSFM, w/o MFIM. (d) w/ FSFM, w/ MFIM in Decoder (Ours). (e) GT.



(a) (b) (c)

Figure 8: Visualization of supplement feature maps from MFIMs in the first branch of the decoder. (a) Supplement feature maps from the fourth branch. (b) Supplement feature maps from the third branch. (c) Supplement feature maps from the second branch.



(a) SRIE (b) DLN (c) ZERO-DCE

(d) EG (e) Ours

Figure 9: Pedestrian detection in the enhancement results of different comparison methods in the dark environment.

ent features, which can supplement the features of each scale.

Effect of the number of network branches. To evaluate the effect of the number of network branches on network performance, we change the number of network branches and test the deformed models on the LOL testset, and the results are shown in Table 5. As can be seen from the table, the model with four branches performs best compared to the models with three branches and five branches. This shows that increasing the depth of the network does not necessarily improve the performance of the network.

Methods	PSNR↑	SSIM↑
Three branches	23.771	0.905
Five branches	24.298	0.903
Four branches (Ours)	**24.732**	**0.912**

Table 5: The effect of the number of network branches.

Therefore, this paper selects a model with four branches as the final network.

Pedestrian Detection in Dark Environments

Image enhancement tasks are often used to serve high-level vision tasks such as object detection, and good enhancement results can improve the accuracy of object detection. Therefore, to verify that the proposed method can better assist high-level vision tasks, several enhancement methods are selected to enhance the test images in the DARK-FACE dataset (Yang et al. 2020) and perform pedestrian detection on the enhanced results, as shown in Figure 9. Here, YOLOv5 is selected for pedestrian detection. As can be observed from the figure that more pedestrians are detected in our result and our result obtains higher detection confidence, which also indicates that our method can obtain better enhancement results and can better help recognize pedestrians.

Conclusion

In this paper, we propose a LIEN-MFC that utilizes a U-shaped encoder-decoder structure to extract and restore image features at different scales. The encoder performs feature extraction by receiving low-light images at four different scales. In the decoder, a FSFM is proposed to achieve complementation and fusion of features at different scales, and an FRM is constructed to obtain the restored features at each scale. To establish supervision for each scale branch, an IRM is constructed to output the enhanced results at each scale, which is used to define the loss function. To better supervise the training of the network, a joint loss function with multiple loss terms is defined. In the loss function, a saturation loss term and a discriminant loss term are designed to make the visual perception of the enhanced image more natural. Experimental results on multiple datasets show that the proposed LIEN-MFC outperforms some state-of-the-art methods.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62072218, No. 61862030, and 62261025).

References

Afifi, M.; Derpanis, K. G.; Ommer, B.; and Brown, M. S. 2021. Learning Multi-Scale Photo Exposure Correction. In *Proceed-ings of the IEEE/CVF Conference on Computer Vision and Pat-tern Recognition (CVPR)*, 9153-63. doi.org/ 10.1109/CVPR46437.2021.00904.

Demir, U., and Unal, G. 2018. Patch-Based Image Inpaint-ing with Generative Adversarial Networks. arXiv:1803.07422.

Dong, X.; Wang, G.; Pang, Y.; Li, W.; Wen, J.; Meng, W.; and Lu, Y. 2011. Fast Efficient Algorithm for Enhance-ment of Low Lighting Video. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 1-6. doi.org/ 10.1109/ICME.2011.6012107.

Fu, X.; Zeng, D.; Huang, Y.; Y. Liao; Ding, X.; and Pais-ley, J. 2016a. A Fusion-Based Enhancing Method for Weakly Illuminated Images. *Signal Processing* 129: 82-96. doi.org/ 10.1016/j.sigpro.2016.05.031.

Fu, X.; Zeng, D.; Huang, Y.; Zhang, X. P.; and Ding, X. 2016b. A Weighted Variational Model for Simultaneous Reflectance and Illumination Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2782-90. doi.org/ 10.1109/CVPR.2016.304.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-Reference Deep Curve Estima-tion for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-tern Recognition (CVPR)*, 1777-86. doi.org/ 10.1109/CVPR42600.2020.00185.

Guo, X.; Li, Y.; and Ling, H. 2017. Lime: Low-Light Im-age Enhancement Via Illumination Map Estimation. *IEEE Transactions on Image Processing* 26(2): 982-93. doi.org/ 10.1109/TIP.2016.2639450.

Hu, J.; Shen, L.; Albanie, S.; Sun, G.; and Wu, E. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(8): 2011-23. doi.org/ 10.1109/TPAMI.2019.2913372.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. Enlightengan: Deep Light Enhancement without Paired Supervision. *IEEE Transactions on Image Processing* 30: 2340-49. doi.org/ 10.1109/TIP.2021.3051462.

Lv, F.; Liu, B.; and Lu, F. 2020. Fast Enhancement for Non-Uniform Illumination Images Using Light-Weight CNNs. In *Proceedings of the 28th ACM International Con-ference on Multimedia (ACM MM)*, 1450–58. doi.org/ 10.1145/3394171.3413925.

Wang, L.; Liu, Z.; Siu, W.; and Lun, D. P. K. 2020. Light-ening Network for Low-Light Image Enhancement. *IEEE Transactions on Image Processing* 29: 7984-96. doi.org/ 10.1109/TIP.2020.3008396.

Wang, S.; Zheng, J.; Hu, H. M.; and Li, B. 2013. Natural-ness Preserved Enhancement Algorithm for Non-Uniform Illumination Images. *IEEE Transactions on Image Pro-cessing* 22(9): 3538-48. doi.org/ 10.1109/TIP.2013.2261309.

Wang, W.; Wei, C.; Yang, W.; and Liu, J. 2018. Gladnet: Low-Light Enhancement Network with Global Awareness. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 751-55. doi.org/ 10.1109/FG.2018.00118.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Pro-cessing* 13(4): 600-12. doi.org/ 10.1109/TIP.2003.819861.

Woo, S.; Park, J.; Lee, J. Y.; and Kweon, I. S. 2018. Cbam: Convolutional Block Attention Module. In *Proceedings of the European conference on computer vision (ECCV)*, 3-19.

Wu, W.; Weng, J.; Zhang, P.; Wang, X.; Yang, W.; and Jiang, J. 2022. Uretinex-Net: Retinex-Based Deep Unfold-ing Network for Low-Light Image Enhancement. In *Pro-ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5901-10.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2021. Band Representation-Based Semi-Supervised Low-Light Image Enhancement: Bridging the Gap between Signal Fi-delity and Perceptual Quality. *IEEE Transactions on Image Processing* 30: 3461-73. doi.org/ 10.1109/TIP.2021.3062184.

Yang, W.; Yuan, Y.; Ren, W.; Liu, J.; Scheirer, W. J.; Wang, Z.; Zhang, T., et al. 2020. Advancing Image Under-standing in Poor Visibility Environments: A Collective Benchmark Study. *IEEE Transactions on Image Pro-cessing* 29: 5737-52. doi.org/ 10.1109/TIP.2020.2981922.

Ying, Z.; Li, G.; and Gao, W. 2017. A Bio-Inspired Multi-Exposure Fusion Framework for Low-Light Image En-hancement. arXiv:1711.00591.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 586-95. doi.org/ 10.1109/CVPR.2018.00068.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond Brightening Low-Light Images. *International Journal of Computer Vision* 129(4): 1013-37. doi.org/ 10.1007/s11263-020-01407-x.

Zhang, Y.; Zhang, J.; and Guo, X. 2019. Kindling the Darkness: A Practical Low-Light Image Enhancer. In *Pro-ceedings of the 27th ACM International Conference on Multimedia (ACM MM)*, 1632–40. doi.org/ 10.1145/3343031.3350926.

Zhao, Z.; Xiong, B.; Wang, L.; Ou, Q.; Yu, L.; and Kuang, F. 2022. Retinexdip: A Unified Deep Framework for Low-Light Image Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* 32(3): 1076-88. doi.org/ 10.1109/TCSVT.2021.3073371.