Deep Symmetric Network for Underexposed Image Enhancement with Recurrent Attentional Learning

Lin Zhao[1*] Shao-Ping Lu[1*] Tao Chen[2] Zhenglu Yang[1] Ariel Shamir[3]
[1]TKLNDST, CS, Nankai University, Tianjin, China
[2]Elephant Technologies, China
[3]The Interdisciplinary Center, Herzliya, Israel
lin-zhao@mail.nankai.edu.cn; {slu,yangzl}@nankai.edu.cn; tao.chen1@vcg.com; arik@idc.ac.il

Abstract

Underexposed image enhancement is of importance in many research domains. In this paper, we take this problem as image feature transformation between the underexposed image and its paired enhanced version, and we propose a deep symmetric network for the issue. Our symmetric network adapts invertible neural networks (INN) for bidirectional feature learning between images, and to ensure the mutual propagation invertible we specifically construct two pairs of encoder-decoder with the same pre-trained parameters. This invertible mechanism with bidirectional feature transformations enable us to both avoid colour bias and recover the content effectively for image enhancement. In addition, we propose a new recurrent residual-attention module (RRAM), where the recurrent learning network is designed to gradually perform the desired colour adjustments. Ablation experiments are executed to show the role of each component of our new architecture. We conduct a large number of experiments on two datasets to demonstrate that our method achieves the state-of-the-art effect in underexposed image enhancement. Code is available at https://www.shaopinglu.net/proj-iccv21/ImageEnhancement.html.

1. Introduction

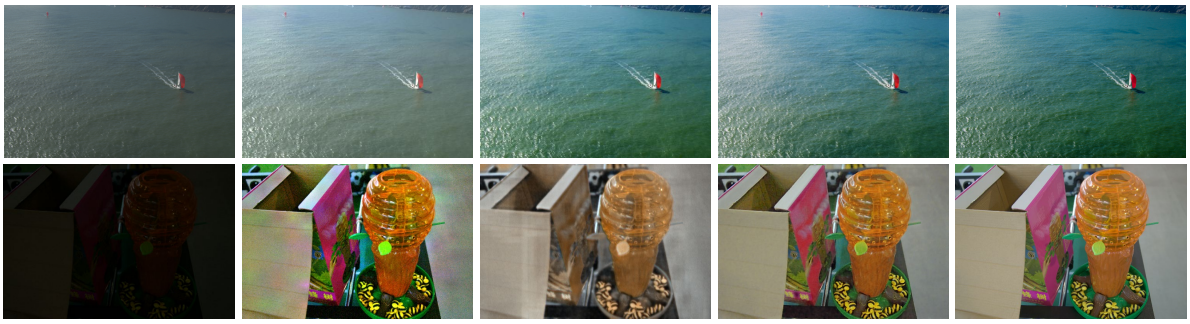
Digital photography is gaining increasing popularity thanks to the abundance of digital cameras widely used in day-to-day life. Still, poor shooting environment, inappropriate camera parameters, or lack of photographic skills can result in unsatisfactory image quality. Many times it is necessary to adjust the exposure-aware aspects of the photograph including the colour and local details in post-processing.

Image enhancement still remains a challenge especially for underexposed images. To enhance the image quality, both colour adjustments are required, and preservation of the content-features of the image. Traditional algorithms with global adjustments, such as histogram equalization [31,40,41], contrast adjustment [18,44] and Gamma correction [22,42] are incapable of editing and changing the local details in the image. Recent methods based on deep neural networks [7, 11, 21, 34, 46] can still suffer from either colour bias or artifacts in complex underexposed conditions. Specifically, when the picture is taken in a low-light environment, the visual features are hidden in dark areas. To correct this situation, not only colour adjustments are needed, but also content recovery (see an example in Fig. 1). Some methods attempt to deal with these requirements by employing multiple different modules separately. However, this scheme may introduce accumulated training errors, and result in visual artifacts.

In this paper, we formulate the image enhancement problem as an unified framework of invertible feature transformation between an image pair: an underexposed image and its enhanced version (that can be the ground truth during training). Therefore, we propose a deep symmetric network based on an invertible feature transformer (IFT) inspired by the latest invertible neural networks (INN) [1, 8, 27, 48]. In order to make the forward and backward propagation operations highly solvable, two pairs of pre-trained encoder-decoder, which exactly share the same parameters, are specifically designed to apply the mutual conversion between the image pair (*i.e.* , the underexposed and enhanced images) and the corresponding features. Our symmetric network carries out the forward and backward learning synchronously, and successfully solves the image colour bias problem caused by the lack of massive training data of paired images and the difficulty of learning color features.

To accurately restore the desired features of the image, we further propose a recurrent learning schedule with a re-

*indicates equal contribution.



(a) Input (b) DeepUPE (c) DeepLPF (d) Ours (e) GT

Figure 1. The results of different methods on challenging images. DeepUPE [46] and DeepLPF [34] are the state-of-the-art methods now. Our method can effectively adjust the image colour while ensuring the structure and texture features of the images.

current residual-attention module (RRAM). Because different weights are assigned to different channels of the feature map, our network can focus on recovering the structure feature information. Without increasing the number of network parameters, the recurrent learning allows our network to learn the complex feature transformation in a stepwise manner, and then realize the adjustments of the image colour. Extensive experiments on public datasets confirm the superiority of our method.

In summary, our contributions can be listed as follows:

- To our best knowledge, we are the first to introduce invertible neural networks (INN) into underexposed image enhancement. Our symmetric architecture performs bidirectional feature learning synchronously, achieving state-of-the-art results compared against other underexposed image enhancement solutions.
- We propose a recurrent learning scheme of feature transformation with a recurrent residual-attention module (RRAM), allowing to apply colour adjustment gradually without increasing network parameters.

2. Related Work

Many research approaches to image enhancement have been introduced in the last decades. Here we briefly discuss some important works especially for underexposed image enhancement. Besides, some most relevant methods to our work will be referred in this section.

Image enhancement. There are many algorithms to adjust the pixel values for image enhancement. Some traditional algorithms propose to enhance the contrast and brightness of the image. For instance, Ying *et al.* [51] use exposure fusion framework to enhance image contrast. Aubry *et al.* [2] employ fast local Laplacian filters to enhance details. Recently, deep learning has been successfully introduced into image enhancement. Gharbi *et al.* [11] introduce a bilateral grid processing network for colour transformation. In order to estimate the global priors and get satisfactory performance, He *et al.* [16] propose a condition network besides the base network. In addition, deep reinforcement learning based methods are exploited to beautify images [21,37]. Some recent work [7,23,24] also build upon generative adversarial networks (GANs) to tackle the problem.

Specifically, many methods focus on enhancing extremely underexposed images. Because the raw images easily lose content information, various methods are proposed to recover such content. Some of these methods also introduce restoring modules to the image enhancement pipeline. However, this class of processing mechanisms would cause accumulated training errors. Chen *et al.* [5] propose an end-to-end pipeline to avoid the training errors, focusing on the raw sensor data instead of low-light RGB images. Xu *et al.* [49] restore the low-frequency and high-frequency layers in turn, and objects are recovered in the low-frequency layer. Besides, according to Retinex theory [28], many researchers take low-light image enhancement as an illumination estimation task [14, 29, 30, 36, 47]. Recently, some specific datasets are also introduced to adapt new training strategies for various image enhancement tasks. For example, Jiang *et al.* [25] adopt an U-Net based GAN for low-light images from different dataset domains. Guo *et al.* [13] present a network that can be trained without ground truth to avoid the risk of overfitting. Yang *et al.* [50] propose DRBN that uses semi-supervised learning to enhance images with perceptual guidance from high-quality images.

Some existing methods attempt to enhance the image while ensuring the feature distribution of the raw image. As an example, Wang *et al.* [46] design an image-to-illumination mapping model to learn the complex image adjustment process. However, there is no global adjustments to the image. Moran *et al.* [34] provide a novel approach that learns spatially local filters to enhance images. However, it does not consider the global filters, and the results might introduce some colour bias. On the contrary, our symmetric network both maintains colour consistency and recovers the content when performing image enhancement.

Finally, there are some other tasks such as face enhancement [10, 39] and shadow enhancement [52]. Compared to

these specific topics, our work is more general for enhancing various underexposed images.

Recurrent attention model. Unlike other feed-forward neural networks, the recurrent neural network (RNN) usually takes the sequential data as input, and it performs with a recursive style in the evolution direction of the sequence [17]. RNN was originally used to solve natural language processing problems [32,45], and recently it has been introduced into computer vision tasks [9,12,35]. Moreover, the recurrent attention was proposed by Mnih *et al.* [33] for image classification. Subsequently, different models with similar ideas are widely used in other tasks. For instance, Chen *et al.* [6] propose a RNN-based visual attention model to learn a sequence of views for 3D shape classification; Haque *et al.* [15] introduce the recurrent attention model into person identification; in [26] a soft attention mechanism is used for object tracking, and Bendre *et al.* [3] put a sequence of frames to the RNN for human action recognition [3]. To our best knowledge, we are the first to introduce the recurrent attention scheme to the underexposed image enhancement problem.

3. Symmetric Network

The overall architecture of our symmetric network is shown in Fig. 2, where two-way propagation operations are conduced during the training process. When the network performs the forward operation (see the green dashed arrows), the input image to be enhanced is processed with the forward transformation. When the backward operation is handled (see the red dashed arrows), the ground truth is reversibly transferred to its underexposed version. This symmetric framework makes our task highly solvable under a bidirectional propagation style.

Our symmetric network contains an invertible feature transformer (IFT), which is based on the latest INNs [1, 8, 27, 48]. Note that due to the lack of depth features, desired image enhancement results could not be obtained when directly applying existing INN architectures on image patches (further details will be given in the ablation section). This calls for more complicated reorganization of image features to meet the hard constraint that the two-way propagation operations should be highly invertible. Therefore, we specifically design two pairs of pre-trained encoder-decoder networks sharing the same parameters. In both propagation sides of our system, the encoder is utilized to perform conversion from images to their corresponding features, while the decoder is designed to transform the features to the corresponding images.

For the forward propagation, we use the first pair of encoder-decoder to convert between images and their corresponding features, and our IFT performs forward feature learning. Formally, this propagation works as:

$$\begin{aligned} x_{f_{L_1}} &= E_1(x_{LQ}), \\ x_{f_{H_1}} &= IFT(x_{f_{L_1}}), \\ x_{HQ_f} &= D_1(x_{f_{H_1}}), \end{aligned} \tag{1}$$

where x_{LQ} represents the input image. $x_{f_{L_1}}, x_{HQ_f}$ and $x_{f_{H_1}}$ denote the output results of the encoder, decoder and IFT, respectively. $[E_1(\cdot),\ D_1(\cdot)]$ and $IFT(\cdot)$ are the forward encoder-decoder and forward feature transformation, respectively. Similarly, the second pair of encoder-decoder and the IFT are involved in the backward propagation, which can be formulated as:

$$\begin{aligned} x_{f_{H_2}} &= E_2(x_{HQ}), \\ x_{f_{L_2}} &= IFT_R(x_{f_{H_2}}), \\ x_{LQ_f} &= D_2(x_{f_{L_2}}), \end{aligned} \tag{2}$$

where x_{HQ} denotes the ground truth. $x_{f_{H_2}}, x_{LQ_f}$ and $x_{f_{L_2}}$ are the corresponding results, respectively. $[E_2(\cdot),\ D_2(\cdot)]$ and $IFT_R(\cdot)$ represent the backward encoder-decoder and backward feature transformation, respectively. Therefore, our network preserves the consistency of the features in both propagation directions, and the two-way constraint solves the colour bias issue for underexposed images.

3.1. Pre-trained Encoder and Decoder

As shown in Fig. 2, we adopt a symmetric encoder-decoder structure for the conversion between images and their features to ensure the integrity of the involved features. Our purpose is to make sure that the pairs of ($x_{f_{L_1}}$, $x_{f_{L_2}}$) and ($x_{f_{H_1}}, x_{f_{H_2}}$) are highly consistent, and that they preserve the global image features, such that the structural information of the image will be well preserved in the underexposed image enhancement.

In order to make sure that the proposed framework invertible, we further constrain that the parameters of the encoder are exactly the same as that of the decoder. In our solution, these parameters are extracted from the first two convolutional layers of a VGG-16 model pre-trained on ImageNet [43]. Note that with more CNN layers, the global features would be damaged, and the reconstruction results would thus be affected.

3.2. Invertible Feature Transformer (IFT)

Underexposed images usually have the loss issue of colour and content. To restore them correctly without artifacts, our IFT learns not only the forward mapping between the low-quality to high-quality image, but also the backward mapping from high to low. Our IFT consists of several invertible blocks (8 by default). For the i-th block, the input feature ${x_f}^i$ is equally divided into ${x_f}_1^i$ and ${x_f}_2^i$ according

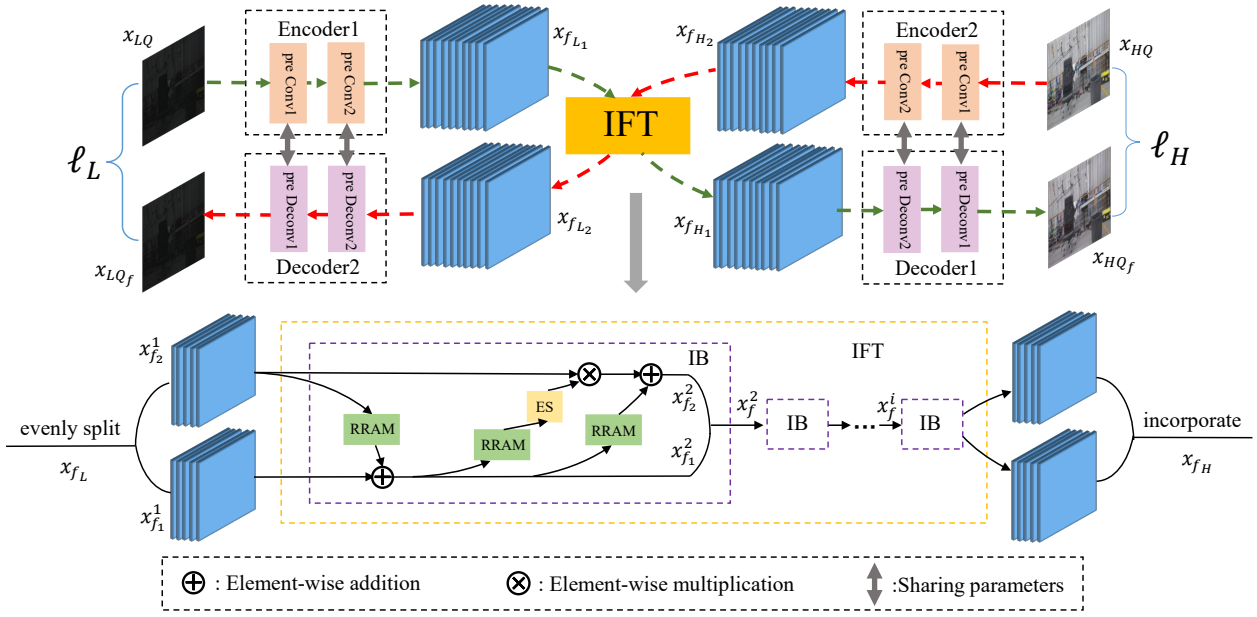


Figure 2. Overall network architecture of our method. The green and red dashed arrows respectively represent the two-way propagation operations of the network. preConv$i$ and preDeconv$i$ refer to the convolution and deconvolution using the i (i=[1, 2]) convolution layer parameters of the pre-trained VGG-16 [43], respectively. Before the feature ${x_f}^i$ enters the i-th invertible block (IB), ${x_f}^i$ is equally divided into ${x_f}_1^i$ and ${x_f}_2^i$ according to the number of channels.

to the channel number, and then passes through the transformation modules:

$$\begin{aligned} {x_f}_1^{i+1} &= \frac{{x_f}_1^i - T_{i,3}({x_f}_2^i)}{ES(T_{i,1}({x_f}_2^i))}, \\ {x_f}_2^{i+1} &= {x_f}_2^i - T_{i,2}({x_f}_1^{i+1}), \end{aligned} \tag{3}$$

where $T_{i,j}(\cdot)$ refers to the j-th (j=1, 2, or 3) transformation module in the i-th block. $ES(\cdot)$ represents the sigmoid function followed by the exponent. We use $ES(\cdot)$ as a multiplier to strengthen the transformation ability. Again, when the backward propagation happens, it is easy to draw that:

$$\begin{aligned} {x_f}_2^i &= {x_f}_2^{i+1} + T_{i,2}({x_f}_1^{i+1}), \\ {x_f}_1^i &= {x_f}_1^{i+1} * ES(T_{i,1}({x_f}_2^i)) + T_{i,3}({x_f}_2^i). \end{aligned} \tag{4}$$

3.3. Recurrent Residual-attention Module (RRAM)

In many pairs of the underexposed and ground truth images, there are obvious colour differences that are very difficult to overcome using existing methods. We thus introduce a recurrent residual-attention module (RRAM) to solve this problem. Considering that transformation learning of colour feature between images is still very challenging, here we adopt the RRAM as a multi-round recurrence module, that learns the target gradually by using the attention architecture for t rounds recurrence without increasing network parameters.

Multi-round recurrent learning. The core idea of our recurrent learning is to divide the task to be solved into several sequential steps, and we learn color adjustment gradually in the task of underexposed image enhancement. As Fig. 3 shows, we use $[h^1, h^2, ..., h^t]$ to represent a sequence of hidden states as an RNN, t is the number of recurrent round. Moreover, the hidden state uses multiple rounds of memory, which can store the feature information obtained in the previous rounds. The features in our network use element-wise addition to alleviate the learning task of each module. Formally, for the t-th round:

$$h^t = \begin{cases} 0, & \text{if } t = 1 \\ h^{t-1} + x_{fo}^{t-1}, & \text{otherwise} \end{cases}, \tag{5}$$

where x_{fo}^{t-1} is the output of the residual soft channel attention mechanism in $(t-1)$-th round. Therefore, a t-th round recurrence is represented as:

$$x_{fo}^t = f_{RSCA}(W_h h^t + W_x x_{fi}), \tag{6}$$

where RSCA presents the residual soft channel attention mechanism, W_h and W_x denote the balanced weights. We select $W_h = W_x = 1$ and $t = 3$. With this design, the hidden state effectively contains the feature information learned previously, and it can focuses on the remaining information missed in each round. Therefore, the RRAM model can efficiently learn the obvious colour difference between the underexposed image and its ground truth.

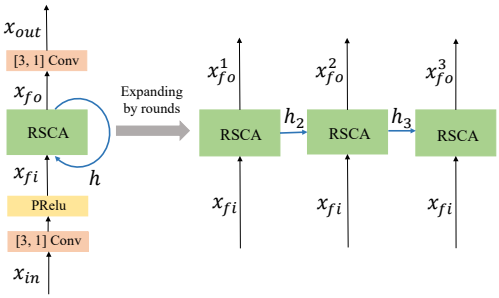


Figure 3. The structure diagram of our RRAM. The blue line represents the transfer of the hidden state. The RSCA means the residual soft channel attention mechanism. The $[i, j]$Conv denotes a convolution operator with $i \times i$ kernel size and $j \times j$ stride size.

Residual soft channel attention mechanism. The residual soft channel attention mechanism is applied to make the network concentrate more on channel-wise structure information. As shown in Fig. 4, the mechanism firstly conducts feature learning in the transformation as follows:

$$x_h = f_C(\delta(f_C(x_{fi}))), \tag{7}$$

where $f_C(\cdot)$ and $\delta(\cdot)$ denote the function of convolution and Prelu, x_{fi} and x_h mean the input of the mechanism and the obtained feature. Next, the inter-dependencies between channels of x_h are dynamically learned. To achieve it, the global average pooling turns the $C \times H \times W$ feature x_h to the $C \times 1 \times 1$ unlearned map M_i as SENet [19], which has a global receptive field to some extent. By learning the attention map, we automatically obtain the importance of each feature channel:

$$M_o = f_{CS}(f_{CR}(M_i)), \tag{8}$$

where $f_{CS}(\cdot)$ and $f_{CR}(\cdot)$ express a convolution followed by sigmoid layer and a convolution followed by Prelu layer. M_o is the obtained soft attention map used to multiply with x_h. In addition, we use residual learning to optimize the network as follows:

$$x_{fo} = x_h \times M_o + x_{fi}, \tag{9}$$

where x_{fo} represents the gained feature of the mechanism.

As mentioned before, the feature is equally divided into two parts when transferred into each invertible block (*i.e.* , the purple dashed box in Fig. 2). To fully extract the feature information of channels and the relationship between such channels, we double and restore the number of feature channels in both the front and end of RRAM,

$$x_{fi} = \delta(f_C(x_{in})), \tag{10}$$

$$x_{out} = f_C(x_{fo}^t), \tag{11}$$

where x_{in} and x_{out} indicate the input and output features of RRAM. x_{fi} and x_{fo}^t are the input and the t-th output of our attention mechanism.

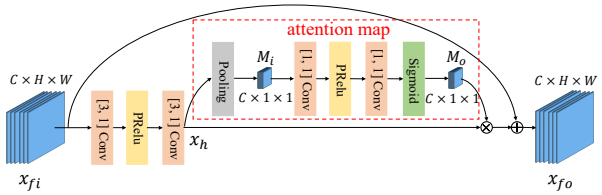


Figure 4. Our residual soft channel attention mechanism. The pooling applies a 2D adaptive average pooling. C, H and W respectively represent the number of channels, length and width of the feature.

3.4. Loss Functions

We use the L_2 distance as the training loss function in both propagation operations. The loss function of the forward propagation operation is defined as:

$$\ell_H = \frac{1}{N}\sum_{i=1}^{N} \left\| x_{HQ_f} - x_{HQ} \right\|^2, \tag{12}$$

where N represents the number of the training images. Similarly, the backward propagation loss function is:

$$\ell_L = \frac{1}{N}\sum_{i=1}^{N} \left\| x_{LQ_f} - x_{LQ} \right\|^2. \tag{13}$$

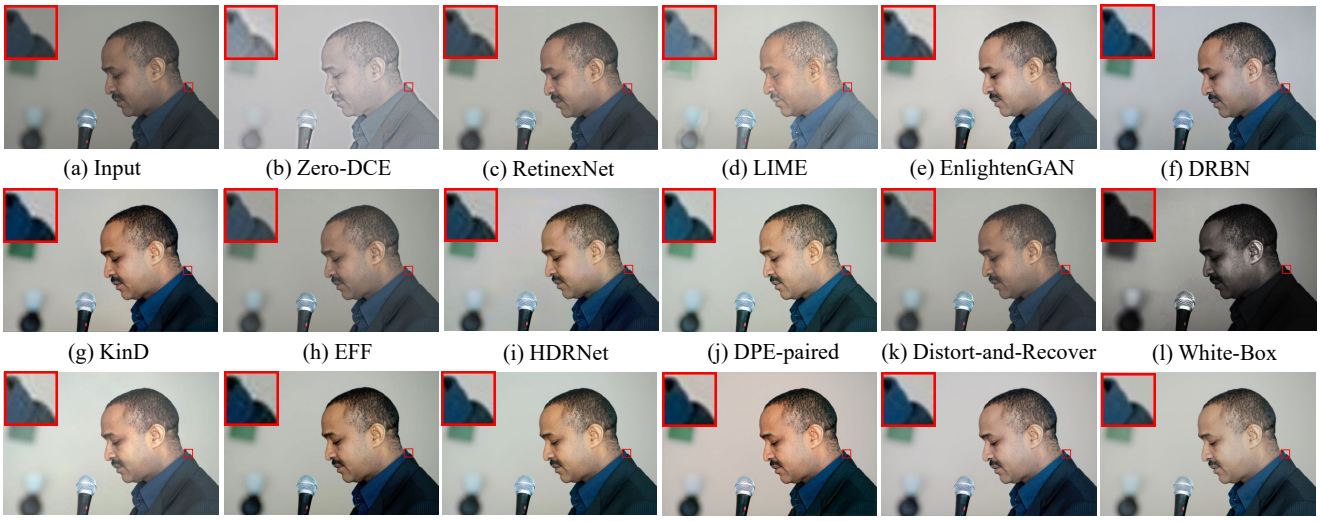
Therefore, the final training loss is defined as a weighted sum of two above-mentioned losses:

$$\ell_{tr} = \lambda_H \ell_H + \lambda_L \ell_L, \tag{14}$$

where λ_H and λ_L refer to the balanced weights. In our training, we empirically set $\lambda_H = \lambda_L = 1$.

4. Experiments

Datasets. We test our network on two benchmark datasets, which are the MIT-Adobe FiveK dataset [4] and LOL dataset [47]. There are 5,000 low-quality images in the MIT-Adobe FiveK dataset, and each image is processed by five experts. We follow [7,34,37,46] to choose the adjusted images of Expert C as ground truth. The original MIT-Adobe FiveK dataset represents only some specific distributions of underexposed images, which may lead to poor generalization when facing large distribution change of the input. In [37] the ground truth is pre-processed with multiple random distorted operators to further synthesize various underexposed results as the input. Here we exactly follow this processing. The first 4,500 images and the last 500 images are used for training and testing respectively. We use LOL dataset [47] to test the performance of our network for extremely underexposed images. LOL dataset [47] contains 500 low/normal-light real image pairs, and we use 400 pairs for training and 100 pairs for testing.



(a) Input (b) Zero-DCE (c) RetinexNet (d) LIME (e) EnlightenGAN (f) DRBN

(g) KinD (h) EFF (i) HDRNet (j) DPE-paired (k) Distort-and-Recover (l) White-Box

(m) DeepUPE (n) DeepLPF (o) CSRNet (p) Ours-without RL (q) Ours (r) GT

Figure 5. The visual results of different methods for the underexposed image in the MIT-Adobe FiveK dataset [4]. Here we also show the result of our method without recurrent learning (RL).

Implementation Details. Our network is implemented on both Pytorch [38] and Jittor [20] platforms with one Nvidia 2080Ti GPU. For the training on each dataset, we randomly crop the original images into 180×180 patches and apply the ADAM optimizer with an initial learning rate of 2e-4. For the MIT-Adobe FiveK dataset, the network is trained for 200 epochs and the learning rate is reduced by half every 50 epochs. For the LOL dataset, the respected epoch numbers are 250 and 100. We utilize PSNR and SSIM as the image quality evaluation standard for the testing. Higher PSNR and SSIM values mean better results.

4.1. Comparison with State-of-the-art

To verify the effectiveness of our method, we compare it with other 14 existing methods: HDRNet [11], DPE (paired) [7], Distort-and-Recover [37], White-Box [21], CSRNet [16] for underexposed images, Zero-DCE [13], RetinexNet [47], LIME [14], EFF [51], EnlightenGAN [25], DRBN [50], KinD [53] for extremely underexposed images, DeepUPE [46], DeepLPF [34] for both. To ensure the fairness of comparison, we retrain all methods.

Quantitative Comparison. We show the results of these methods on the two datasets in Tab. 1. The three sections from top to bottom are the methods that focus on low-light enhancement, image retouching and both. Our method achieves the highest values except SSIM on LOL dataset.

Qualitative Comparison. Evaluations are applied on two datasets. We select an underexposed image from the MIT-Adobe FiveK dataset to show the comparison with all methods in Fig. 5. It can be seen that Zero-DCE, LIME and White-Box have colour bias, RetinexNet, EFF, DRBN and Distort-and-Recover fail to adjust image brightness,

Method	LOL	FiveK
	PSNR/SSIM	PSNR/SSIM
Zero-DCE [13]	13.08/0.470	12.30/0.673
RetinexNet [47]	17.03/0.707	20.20/0.781
LIME [14]	16.92/0.540	14.30/0.731
EnlightenGAN [25]	17.79/0.769	21.28/0.818
EFF [51]	16.94/0.592	18.15/0.784
DRBN [50]	19.24/**0.847**	21.71/0.855
KinD [53]	20.08/0.822	21.72/0.833
HDRNet [11]	19.62/0.716	23.29/0.842
DPE (paired) [7]	18.08/0.659	21.67/0.846
Distort [37]	20.46/0.666	21.29/0.812
White-Box [21]	17.59/0.633	17.30/0.755
CSRNet [16]	19.57/0.681	24.13/0.878
DeepUPE [46]	16.78/0.468	20.83/0.795
DeepLPF [34]	16.58/0.678	23.63/0.875
w/o RL	20.63/0.826	23.32/0.888
ours	**21.71**/0.834	**24.27/0.900**

Table 1. Quantitative results of different methods on the MIT-Adobe FiveK dataset [4] and LOL dataset [47]. "w/o RL" refers to our method without recurrent learning and "Distort" is Distort-and-Recover.

DeepLPF does not recover the local information. DPE-paired and CSRNet do not adjust the image colour correctly. EnlightenGAN, KinD, HDRNet, Distort-and-Recover and DeepUPE produce artifacts in images that cause some texture details to disappear.

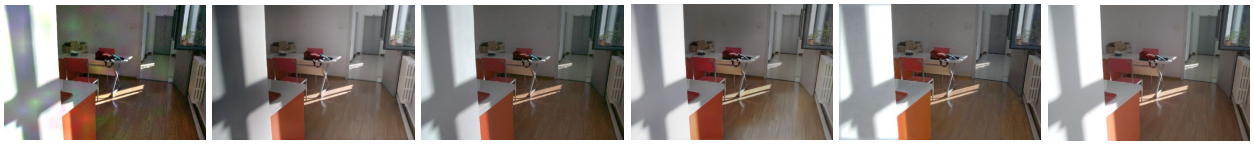
As shown in Fig. 6, we compare with the methods for image retouching to show our strength in this respect. When there is colour difference between the input and ground truth image pair, our result is more consistent with the

(a) Input (b) HDRNet (c) DPE-paired (d) Distort-and-Recover (e) White-Box

(f) DeepUPE (g) DeepLPF (h) CSRNet (i) Ours (j) GT

Figure 6. Comparison results with image retouching methods on the image in the MIT-Adobe FiveK dataset [4].

(a) Input (b) Zero-DCE (c) RetinexNet (d) LIME (e) EnlightenGAN (f) DeepLPF



(g) DeepUPE (h) DRBN (i) EFF (j) KinD (k) Ours (l) GT

Figure 7. The results of low-light image enhancement methods on the image in LOL dataset [47].

ground truth than other methods.

In addition, we compare with the low-light image enhancement methods on the LOL dataset, and Fig. 7 shows the results of some extremely underexposed scenes. In general, those results obtained from RetinexNet, DeepLPF, KinD, and DRBN do not restore colour well. Moreover, it reveals that some methods such as Zero-DCE, LIME, DeepUPE, EFF and EnlightenGAN are more likely to generate visual artifacts. Overall, our method achieves better results for enhancing extremely underexposed images.

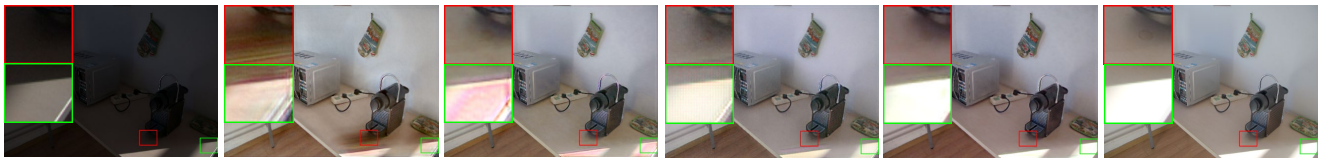
4.2. Ablation Study

As shown in the Fig. 7 and Tab. 1, most methods can not achieve satisfactory results on LOL dataset, so we use the dataset for ablation experiments to explain in detail the role of our each module.

In Tab. 2, we conduct experiments from three different aspects to verify the effect of the two-way transformation and loss, RRAM, and symmetrical architecture. For the two-way transformation, we cancel the backward loss to verify the intention of IFT. We also change the loss function to l1, to maximize PSNR [7] for comparison. For RRAM, we change it to other networks. Furthermore, we remove the attention mechanism to evaluate the effectiveness of it, and change the number of recurrences to justify the recurrent learning. To verify the role of symmetrical architecture, we change the number of invertible blocks, and specially replace our pre-trained encoder-decoder with Haar wavelets that are typically used in the previous INN-based networks [27, 48]. The PSNR and SSIM of each variation in Tab. 2 show that each module contributes to the performance improvement.

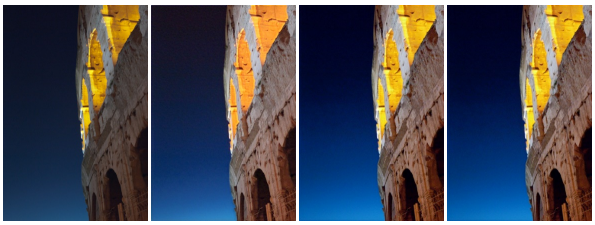
To further illustrate the specific role of attention and symmetric architecture with two-way loss, in Fig. 8 we show the comparison results on some representative images. Fig. 8 (b) indicates that our method fails to recover some content once removing attention in RRAM. When the network does not perform backward learning (see Fig. 8 (c)), it causes errors in the illuminated part of image. Moreover, the results based on Haar wavelets have obvious artifacts.

Lastly, we also verify the effectiveness of our recurrent learning scheme. As can be seen in Fig. 5 (p)-(q) and Fig. 9 (b)-(c), when we compare the results of our network with or without recurrent learning, it is clear that only with recurrent learning, the network can correctly learn the adjustments of image colour. Also, the PSNR/SSIM value will drop (Tab. 1) on both datasets without recurrent learning.



(a) Input (b) without attention (c) with one-way loss (d) Haar wavelets (e) ours (f) GT

Figure 8. Visual results of different ablation experiments on our method for images in LOL dataset [47].



(a) Input (b) Ours-without RL (c) Ours-with RL (d) GT

Figure 9. Visual results with/without our recurrent learning for images that obvious colour adjustments should be performed.

Condition	PSNR	SSIM
$L_2 \rightarrow L_1$ loss	-0.98	-0.024
$L_2 \rightarrow$ PSNR loss	-0.20	-0.007
w/o backward loss	-1.28	-0.014
w/o attention	-3.10	-0.038
RESBLK	-3.50	-0.067
SE-RESBLK	-1.32	-0.034
DenseNet	-2.74	-0.039
w/o RL	-1.08	-0.008
two-round RL	-0.42	-0.009
4 IB	-1.42	-0.026
Haar wavelets	-1.74	-0.027

Table 2. Quantitative comparison of different ablation experiments against the default setups on our method in LOL dataset [47]. Here "RL" refers to recurrent learning, "IB" means invertible block, and "SE-RESBLK" is the residual block embedded with SENet [19].

4.3. User Study

Image retouching performance and aesthetics are strongly affected by individual bias and expertise. To better evaluate our method, we invite twenty participants to perform a user study, wherein five experts are professional photographers/editors with years of experiences working for stock photo agencies, while the others are randomly searched amateurs of the image enhancement domain. We select 50 raw photos from their day-to-day retouching tasks of five categories, namely "People", "Night", "Architecture", "Nature", and "Drastic Weather". For each category, we randomly select 10 photos. We believe this test set is a good representative of a real world retouching task. For each image, we generate four retouching results from CSRNet, DeepLPF, DeepUPE and our method respectively. The three existing methods are picked for their better performance in the quantitative evaluation and manageable workload. All models are the trained models from the MIT-Adobe FiveK dataset to test the robustness of the networks. Each tester has been asked to choose the best in the retouched results of each photo by comprehensively considering their colour, brightness, contrast and artifacts (the photos and results are in the supplementary materials). We separately counted the percentage of each method in the selection results of experts and amateurs. As can be shown in Tab. 3 that no matter the testers are experts or amateurs, our method is preferred in the highest proportion.

Me / Pe	CSRNet	DeepUPE	DeepLPF	ours
experts	23.20%	7.60%	8.80%	**60.40%**
amateurs	24.13%	8.40%	22.13%	**45.33%**

Table 3. The percentage of preferred methods in the user study. 'Me' and 'Pe' present the method and percentage.

5. Conclusions

In the paper, we have proposed a symmetrical deep network, which includes an invertible feature transformer (IFT) and two pairs of pre-trained encoder-decoder. The symmetrical architecture allows to propagate in both directions during training, preserving the feature consistency of the underexposed and the enhanced image pair. In addition, our recurrent residual-attention module (RRAM) helps the system to better achieve desired colour adjustments with complex feature transformation. Moreover, by paying attention to the interdependencies between feature channels, the residual soft channel attention mechanism in RRAM makes our network better restore the structure features. We conducted lots of quantitative and qualitative comparison experiments to prove the superiority of our method.

Acknowledgements. We would like to thank reviewers and ACs for their valuable comments. This work was funded partially by NSFC (No. 61972216 and No. 62111530097), Tianjin NSF (No. 18JCYBJC41300 and No. 18ZXZNGX00110), BNRist (No. BNR2020KF01001) and the Israel Science Foundation grant number 1390/19. Shao-Ping Lu is the corresponding author of this paper.

References

[1] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019. 1, 3

[2] Mathieu Aubry, Sylvain Paris, Samuel W Hasinoff, Jan Kautz, and Frédo Durand. Fast local laplacian filters: Theory and applications. *ACM TOG*, 33(5):1–14, 2014. 2

[3] Nihar Bendre, Nima Ebadi, John J Prevost, and Peyman Najafirad. Human action performance using deep neuro-fuzzy recurrent attention model. *IEEE Access*, 8:57749–57761, 2020. 3

[4] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, pages 97–104, 2011. 5, 6, 7

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, pages 3291–3300, 2018. 2

[6] Songle Chen, Lintao Zheng, Yan Zhang, Zhixin Sun, and Kai Xu. Veram: View-enhanced recurrent attention model for 3d shape classification. *IEEE TVCG*, 25(12):3244–3257, 2018. 3

[7] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *CVPR*, pages 6306–6314, June 2018. 1, 2, 5, 6, 7

[8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 1, 3

[9] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015. 3

[10] Shan Du and Rabab K Ward. Adaptive region-based image enhancement method for robust face recognition under variable illumination conditions. *IEEE TCSVT*, 20(9):1165–1175, 2010. 2

[11] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM TOG*, 36(4):1–12, 2017. 1, 2, 6

[12] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 3

[13] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, pages 1780–1789, 2020. 2, 6

[14] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016. 2, 6

[15] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, pages 1229–1238, 2016. 3

[16] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *ECCV*, 2020. 2, 6

[17] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. 3

[18] A Howie. Image contrast and localized signal selection techniques. *Journal of Microscopy*, 117(1):11–23, 1979. 1

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 5, 8

[20] Shi-Min Hu, Dun Liang, Guo-Ye Yang, Guo-Wei Yang, and Wen-Yang Zhou. Jittor: a novel deep learning framework with meta-operators and unified graph execution. *Science China Information Sciences*, 63(12):1–21, 2020. 6

[21] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM TOG*, 37(2):1–17, 2018. 1, 2, 6

[22] Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE TIP*, 22(3):1032–1041, 2012. 1

[23] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*, pages 3277–3285, 2017. 2

[24] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Wespe: Weakly supervised photo enhancer for digital cameras. In *CVPRW*, pages 691–700, 2018. 2

[25] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972*, 2019. 2, 6

[26] Samira Ebrahimi Kahou, Vincent Michalski, Roland Memisevic, Christopher Pal, and Pascal Vincent. Ratm: recurrent attentive tracking model. In *CVPRW*, pages 1613–1622, 2017. 3

[27] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, pages 10215–10224, 2018. 1, 3, 7

[28] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 2

[29] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE TIP*, 27(6):2828–2841, 2018. 2

[30] Xujie Li, Hanli Zhao, Guizhi Nie, and Hui Huang. Image recoloring using geodesic distance based color harmonization. *Computational Visual Media*, 1(2):143–155, 2015. 2

[31] Shao-Ping Lu, Sen-Mao Li, Rong Wang, Gauthier Lafruit, Ming-Ming Cheng, and Adrian Munteanu. Low-rank constrained super-resolution for mixed-resolution multiview video. *IEEE TIP*, 30:1072–1085, 2021. 1

[32] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *ICASSP*, pages 5528–5531, 2011. 3

[33] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *arXiv preprint arXiv:1406.6247*, 2014. 3

[34] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *CVPR*, pages 12826–12835, 2020. 1, 2, 5, 6

[35] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE TGRS*, 55(7):3639–3655, 2017. 3

[36] Qi Mu, Xinyue Wang, Yanyan Wei, and Zhanli Li. Low and non-uniform illumination color image enhancement using weighted guided image filtering. *Computational Visual Media*, 2021. 2

[37] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *CVPR*, pages 5928–5936, 2018. 2, 5, 6

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[39] Eli Peli, Estella Lee, Clement L Trempe, and Sheldon Buzney. Image enhancement for the visually impaired: the effects of enhancement on face recognition. *JOSA A*, 11(7):1929–1939, 1994. 2

[40] Stephen M Pizer. Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group. In *Proceedings of the First Conference on Visualization in Biomedical Computing, Atlanta, Georgia*, volume 337, 1990. 1

[41] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987. 1

[42] Shanto Rahman, Md Mostafijur Rahman, Mohammad Abdullah-Al-Wadud, Golam Dastegir Al-Quaderi, and Mohammad Shoyaib. An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, 2016(1):1–13, 2016. 1

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014. 3, 4

[44] J Alex Stark. Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE TIP*, 9(5):889–896, 2000. 1

[45] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NeurIPS*, pages 2440–2448, 2015. 3

[46] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, pages 6849–6857, 2019. 1, 2, 5, 6

[47] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *BMVC*, 2018. 2, 5, 6, 7, 8

[48] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. *arXiv preprint arXiv:2005.05650*, 2020. 1, 3, 7

[49] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*, pages 2281–2290, 2020. 2

[50] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, pages 3063–3072, 2020. 2, 6

[51] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new image contrast enhancement algorithm using exposure fusion framework. In *CAIP*, pages 36–46, 2017. 2, 6

[52] Xuaner Cecilia Zhang, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, David E Jacobs, et al. Portrait shadow manipulation. *arXiv preprint arXiv:2005.08925*, 2020. 2

[53] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, pages 1632–1640, 2019. 6