Deep Symmetric Network for Underexposed Image Enhancement with Recurrent Attentional Learning

Lin Zhao¹* Shao-Ping Lu¹* Tao Chen² Zhenglu Yang¹ Ariel Shamir³
¹TKLNDST, CS, Nankai University, Tianjin, China
²Elephant Technologies, China
³The 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.

*indicates equal contribution.

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-
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 step-wise 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 conducted 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:

\[ x_{fL_1} = E_1(x_{LQ}), \]
\[ x_{fH_1} = IFT(x_{fL_1}), \]
\[ x_{HQ_1} = D_1(x_{fH_1}), \]

where \( x_{LQ} \) represents the input image, \( x_{fL_1}, x_{HQ_1} \) and \( x_{fH_1} \) 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:

\[ x_{fH_2} = E_2(x_{HQ}), \]
\[ x_{fL_2} = IFT_R(x_{fH_2}), \]
\[ x_{LQ_2} = D_2(x_{fL_2}), \]

where \( x_{HQ} \) denotes the ground truth. \( x_{fH_2}, x_{LQ_2} \) and \( x_{fL_2} \) 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 \((x_{fL_1}, x_{fL_2})\) and \((x_{fH_1}, x_{fH_2})\) 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 \( x_{fi} \) is equally divided into \( x_{fi1} \) and \( x_{fi2} \) according
to the channel number, and then passes through the transformation modules:

\[
\begin{align*}
    x_{f_1}^{i+1} &= x_{f_1}^i - T_i,3(x_{f_1}^i) \\
    x_{f_2}^{i+1} &= x_{f_2}^i - T_i,2(x_{f_2}^i) + ES(T_{i,1}(x_{f_2}^i))
\end{align*}
\]  

where \(T_i,j(\cdot)\) refers to the \(j\)-th (\(j=1, 2, 3\)) 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:

\[
\begin{align*}
    x_{f_2}^i &= x_{f_2}^{i+1} + T_{i,2}(x_{f_1}^{i+1}) \\
    x_{f_1}^i &= x_{f_1}^{i+1} + ES(T_{i,1}(x_{f_2}^i)) + T_{i,3}(x_{f_2}^i).
\end{align*}
\]

### 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:

\[
h^t = \begin{cases} 
    0, & \text{if } t = 1 \\
    h^{t-1} + x_f^{t-1}, & \text{otherwise} 
\end{cases}
\]

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

\[
x_f^t = f_{RSCA}(W_h h^t + W_x x_f^t),
\]

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.
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:

\[ x_h = f_C(\delta(f_C(x_{f_i}))) \]  

where \( f_C(\cdot) \) and \( \delta(\cdot) \) denote the function of convolution and Prelu, \( x_{f_i} \) 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:

\[ M_o = f_{CS}(f_{CR}(M_i)) \]  

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:

\[ x_{f_o} = x_h \times M_o + x_{f_i} \]  

where \( x_{f_o} \) 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,

\[ x_{f_i} = \delta(f_C(x_{in})) \]  

\[ x_{out} = f_C(x_{f_o}^t) \]  

where \( x_{in} \) and \( x_{out} \) indicate the input and output features of RRAM. \( x_{f_i} \) and \( x_{f_o}^t \) are the input and the \( \ell \)-th output of our attention mechanism.

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:

\[ \ell_H = \frac{1}{N} \sum_{i=1}^{N} \| x_{HQ_i} - x_{HQ} \|^2 \]  

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

\[ \ell_L = \frac{1}{N} \sum_{i=1}^{N} \| x_{LQ_i} - x_{LQ} \|^2 \]  

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

\[ \ell_{tr} = \lambda_H \ell_H + \lambda_L \ell_L \]  

where \( \lambda_H \) and \( \lambda_L \) 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.
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).

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, 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...
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 $l_1$ 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.
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 experience 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, DeepUPE, DeepLPF 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.

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

<table>
<thead>
<tr>
<th></th>
<th>Me</th>
<th>Pe</th>
</tr>
</thead>
<tbody>
<tr>
<td>experts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSRNet</td>
<td>23.20%</td>
<td>7.60%</td>
</tr>
<tr>
<td>DeepUPE</td>
<td>8.80%</td>
<td>60.40%</td>
</tr>
<tr>
<td>DeepLPF</td>
<td>22.13%</td>
<td>45.33%</td>
</tr>
<tr>
<td>amateurs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSRNet</td>
<td>24.13%</td>
<td>8.40%</td>
</tr>
<tr>
<td>DeepUPE</td>
<td>22.13%</td>
<td>45.33%</td>
</tr>
<tr>
<td>DeepLPF</td>
<td>45.33%</td>
<td>45.33%</td>
</tr>
</tbody>
</table>

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].
References


[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


