Empowering Low-Light Image Enhancer through Customized Learnable Priors

Naishan Zheng 1*, Man Zhou 1*, Yanmeng Dong, Xiangyu Rui 2, Jie Huang 1, Chongyi Li 3, Feng Zhao 1†
1University of Science and Technology of China
2Xi’an Jiaotong University, 3Nankai University
{nzheng,manman,hj0117}@mail.ustc.edu.cn,
{dongym.aca,xyrui.aca,lichongyi25}@gmail.com, fzhao956@ustc.edu.cn

Abstract

Deep neural networks have achieved remarkable progress in enhancing low-light images by improving their brightness and eliminating noise. However, most existing methods construct end-to-end mapping networks heuristically, neglecting the intrinsic prior of image enhancement task and lacking transparency and interpretability. Although some unfolding solutions have been proposed to relive these issues, they rely on proximal operator networks that deliver ambiguous and implicit priors. In this work, we propose a paradigm for low-light image enhancement that explores the potential of customized learnable priors to improve the transparency of the deep unfolding paradigm. Motivated by the powerful feature representation capability of Masked Autoencoder (MAE), we customize MAE-based illumination and noise priors and redevelop them from two perspectives: 1) structure flow: we train the MAE from a normal-light image to its illumination properties and then embed it into the proximal operator design of the unfolding architecture; and 2) optimization flow: we train MAE from a normal-light image to its gradient representation and then employ it as a regularization term to constrain noise in the model output. These designs improve the interpretability and representation capability of the model. Extensive experiments on multiple low-light image enhancement datasets demonstrate the superiority of our proposed paradigm over state-of-the-art methods. Code is available at https://github.com/zheng980629/CUE.

1. Introduction

Low-light conditions often result in images with limited visibility and noise, which can negatively affect downstream computer vision tasks. To recover details buried in low-light images and remove noise effects, the field of low-light image enhancement has received significant attention. Existing approaches can be classified into two categories: traditional methods and deep learning-based methods.

Traditional methods for low-light image enhancement involve formulating the problem as an optimization task and using image priors as regularization terms to constrain the solution space. One such representative method is based on the Retinex theory[23], which assumes that an image can be decomposed into reflectance and illumination components where the former keeps consistency under any lighting conditions and the latter reflects variations in bright-
ness. Nonetheless, estimating illumination and reflectance terms simultaneously is challenging. To overcome this issue, based on the Retinex theory, well-designed priors of the illumination and reflectance terms are proposed. For example, an $\ell_2$ norm on illumination gradients is proposed by [12, 11] to ensure smoothness but generate blurred borders around areas where the illumination suddenly changes. [28] further introduces a noise term and develops a $\ell_1$ prior to constrain illumination gradients, maintaining the overall structure of the illumination map. However, hand-crafted priors are difficult to design and have limited representative ability in complex scenes, hindering their practical usage.

Inspired by the powerful learning capability of deep learning, explosive learning-driven methods have undeniably ushered in a new era for low-light image enhancement [18, 17, 20, 57, 19, 16, 63, 64, 31, 39, 33, 38, 42, 22, 53, 37, 41, 30]. For example, RetinexNet [50] and KinD [60] are two popular methods that use deep learning to enhance low-light images. RetinexNet integrates Retinex decomposition and illumination adjustment into an end-to-end trainable network, while KinD separately trains layer decomposition, illumination adjustment, and reflectance restoration subnetworks. Despite their success, these methods often construct black-box networks without considering the intrinsic priors of illumination and reflectance components, leading to a lack of transparency and interpretability.

To improve the interpretability, a model-driven deep unfolding paradigm, URetinexNet [52], has been proposed. It formulates the Retinex decomposition as an implicit prior regularized model but neglects the effects of noise pollution. Moreover, this method delivers the illumination and reflectance priors in a vague and ambiguous manner through empirically constructed proximal operator networks. Thus, the inherent properties of these two components are not considered, leading to an ambiguous and implicit prior principle (see Fig. 1). We therefore wonder, “Can we customize learnable priors for illumination and noise terms that leverage their intrinsic properties?”.

To answer this question, we first analyze the characteristics of the illumination and noise components: 1) illumination prior: Based on the Retinex theory, the illumination component should be smooth and preserve the structure while adapting to different lighting conditions; 2) noise prior: In low-light images, noise is inherent and cannot be removed simply by adjusting brightness, i.e., irrelevant to enhanced lightness.

Motivated by the above analysis, we aim at exploring the potential of customized learnable priors for low-light image enhancement to improve the transparency of the deep unfolding paradigm. Our proposed method is called Customized Unfolding Enhancer (CUE). To achieve this, we utilize the innate feature representation capability of Masked Autoencoder (MAE) to customize MAE-based illumination and noise priors with a masked image modeling strategy. Specifically, the illumination prior is trained from a normal-light image to its illumination map filtered by a bilateral filter, reducing noise without altering the intrinsic structure [6]. The noise prior aims to learn the histograms of oriented gradients of a normal-light image, which presents the gradient variation while being irrelevant to enhanced lightness [8]. To integrate the customized priors into the low-light image enhancement process, we redevelop the two learned priors from two perspectives: 1) structure flow: embedding the learned illumination prior into the design of the proximal operator in the Retinex decomposition unfolding process; and 2) optimization flow: redeveloping the learned noise prior as a regularization term to eliminate noise by minimizing the gradient presentation difference between the enhanced and normal-light images.

Extensive experimental results demonstrate the superiority of our paradigm over state-of-the-art methods. Additionally, we also verify the effectiveness of the proposed learnable noise prior for image denoising.

Our contributions are summarized as follows:

- We activate the potential of customized learnable illumination and noise priors via a new deep unfolding paradigm for low-light image enhancement.
- From the structure flow, we embed the MAE-based customized illumination prior into the unfolding architecture to improve the transparency and interpretability of the unfolding structure.
- From the optimization flow, we redevelop the MAE-based customized noise prior as a regularization term to constrain the gradient representation consistency.
- Our experiments on multiple low-light image benchmarks show that the proposed paradigm outperforms state-of-the-art methods, and our customized noise prior is effective for image denoising.

2. Related Work

2.1. Low-Light Image Enhancement

There are three groups of traditional image enhancement techniques: Histogram Equalization (HE), Gamma Correction (GC), and Retinex theory [23]. HE-based techniques stretch the dynamic range of the low-light images with various complicated priors [21, 3, 5, 55]. GC-based [45, 48] methods expand the dynamic range of an image by applying an exponential function to each pixel individually. In addition, Retinex-based variants [12, 28, 47, 28] have developed priors to constrain the solution space of the illumination and reflectance maps. However, the inflexible hand-crafted priors may not generalize well in complex scenes.
In the past decade, deep learning-based low-light image enhancement approaches [34, 50, 60, 29, 54, 35, 10, 27, 49, 61, 25, 67, 24, 66, 26, 65], have achieved remarkable performance gains. Wei et al. [50] first attempted to integrate Retinex decomposition and illumination adjustment into an end-to-end network. Zhang et al. [60] employed three subnetworks for decomposition, reflection restoration, and illumination adjustment. Wang et al. [46] proposed to enhance the underexposed images by estimating an illumination map. Guo et al. [13] adjusted the dynamic range of the low-light images by predicting pixel-wise mapping without paired or unpaired data. However, these learning-based methods empirically build network architectures in a black-box fashion, neglecting the inherent prior and lacking adequate interpretability. To address this limitation, Liu et al. [32] developed a Retinex-inspired unrolling strategy to discover the architectures embedded with atomic priors automatically. Wu et al. [52] proposed to unroll the Retinex decomposition and formulated it as an implicit prior regularized model. Zheng et al. [62] unfolded the total variational minimization algorithms to provide fidelity and smoothness constraints via a learnable noise level map. While these deep unfolding solutions improve the interpretability, the priors are still vaguely delivered by convolutional neural networks constructed empirically in a black-box manner, resulting in an ambiguous and implicit prior principle.

2.2. Masked Image Modeling

Inspired by the success of BERT [9] in NLP, BEiT [4] proposes a masked image modeling technique that predicts the visual token of each masked patch as a pretext task. MAE [15] presents an asymmetric encoder-decoder paradigm that focuses on the visible patch. The encoder operates on this patch, while the decoder reconstructs the original image from the latent representation and mask tokens. To improve representation capability, CAE [7] decouples the encoding and pretext task completion roles through an alignment constraint and a latent contextual regressor. MVP [51] aims to create robust and multi-modal representation by predicting the guidance from other modalities. However, their primary purpose is to serve as a pre-training technique to enhance performance on downstream tasks.

3. Methodology

In this section, we introduce the proposed Customized Unfolding Enhancer (CUE) paradigm, which consists of a structure flow and an optimization flow, as illustrated in Fig. 2. In the structure flow, we first describe an unfolding Retinex decomposition step embedded with an MAE-based customized illumination prior and provide an enhancement step. Then we go into the optimization flow, including Retinex decomposition, enhancement, and a gradient representation regularization derived from an MAE-based customized noise prior.

3.1. Structure Flow

3.1.1 Retinex Decomposition

Classical Retinex theory assumes that an observed image can be decomposed into illumination and reflectance components. Due to the inevitable noise in low-light images, [28] introduces a noise term:

\[ I = R \odot L + N, \]

where I, R, L, and N denote the observed image, reflectance, illumination, and noise, respectively. The operator \( \odot \) represents the element-wise multiplication. Referring to the above observation model, the reflectance, illumination, and noise terms can be simultaneously obtained by
solving the minimization problem:

$$\arg\min_{\mathbf{R}, \mathbf{L}, \mathbf{N}} \|\mathbf{N}\|_1 + \gamma \rho_1(\mathbf{N}) + \beta \rho_2(\mathbf{L}) + \omega \rho_3(\mathbf{R}), \quad (2a)$$

$$\text{s.t.} \mathbf{I} = \mathbf{R} \circ \mathbf{L} + \mathbf{N}, \quad (2b)$$

where $\| \cdot \|_1$ denotes $\ell_1$ norm, $\rho_1$, $\rho_2$, and $\rho_3$ are the regularization terms denoting imposed priors over $\mathbf{N}$, $\mathbf{L}$, and $\mathbf{R}$, and $\gamma$, $\beta$, and $\omega$ are trade-off parameters. $\|\mathbf{N}\|_1$ is a general prior that simply constrains the noise’s sparsity, which cannot accurately model the noise distribution in the low-light image. Therefore, we incorporate an implicit noise prior, $\rho_1(\mathbf{N})$, to further estimate the extreme noise.

By introducing a penalty function to remove the equalization constraint, Eq. (2) is written as:

$$\mathcal{L}(\mathbf{N}, \mathbf{L}, \mathbf{R}) = \|\mathbf{N}\|_1 + \gamma \rho_1(\mathbf{N}) + \beta \rho_2(\mathbf{L}) + \omega \rho_3(\mathbf{R}) + \frac{\mu}{2} \|\mathbf{I} - \mathbf{R} \circ \mathbf{L} - \mathbf{N}\|_F^2, \quad (3)$$

where $\| \cdot \|_F$ represents Frobenius norm and $\mu$ is a penalty parameter. The equivalent objective function can be solved by iteratively updating $\mathbf{L}$, $\mathbf{R}$, and $\mathbf{N}$ while considering other variables that have been estimated in previous iterations as constants. Here we present the sub-problem solutions for the $(k+1)$-th iteration.

**L sub-problem:** Given the estimated reflectance and noise at iteration $k$, $\mathbf{R}_k$, and $\mathbf{N}_k$, $\mathbf{L}$ can be updated as:

$$\mathbf{L}_{k+1} = \arg\min_{\mathbf{L}} \mathcal{L}(\mathbf{N}_k, \mathbf{L}, \mathbf{R}_k)$$

$$= \arg\min_{\mathbf{L}} \beta \rho_2(\mathbf{L}) + \frac{\mu_k}{2} \|\mathbf{I} - \mathbf{R}_k \circ \mathbf{L} - \mathbf{N}_k\|_F^2.$$

(4)

By applying the proximal gradient method to Eq. (4), we can derive:

$$\mathbf{L}_{k+1} = \text{prox}_{\beta \rho_2}(\mathbf{L}_k - \alpha_1 \nabla g(\mathbf{L}_k)), \quad (5)$$

where $\text{prox}_{\beta \rho_2}$ is the proximal gradient operator corresponding to the prior $\rho_2$, $\alpha_1$ denotes the updating step size, and $\nabla g(\mathbf{L}_k) = \mu_k \mathbf{R}_k \circ (\mathbf{R}_k \circ \mathbf{L}_k + \mathbf{N}_k - \mathbf{I})$.

Existing methods for designing the functional form of $\text{prox}_{\beta \rho_2}$ construct implicit networks in a black box manner, resulting in an ambiguous and implicit prior principle. However, according to Retinex theory, the illumination term should be smooth while preserving the structure and adapting to the diverse lighting conditions [23]. Fortunately, the bilateral filter operator is inherently capable of filtering noise while preserving edges [6]. In this paper, standing on the shoulder of MAE [15], which is trained using the masked image modeling strategy and embraces powerful feature representation capability, we design a pre-trained MAE with the illumination map of a normal-light image filtered by the bilateral filter as the target feature, thereby creating a customized illumination prior. We then embed the learned illumination prior into the design of $\text{prox}_{\beta \rho_2}$ to enhance transparency.

The original MAE is an asymmetric encoder-decoder architecture that reconstructs the masked pixels given partially visible patches. During pre-training, a high proportion of the input image is masked, and the encoder only maps the visible patches into a latent feature representation. Then, a lightweight decoder reconstructs the original images in pixels from the latent representation and mask tokens. After the pre-training, the encoder is discarded, and the decoder is applied to uncorrupted images (full sets of patches) for feature representation.

In this paper, as shown in Fig. 3, deviated from the original MAE, we implement a UNet-like convolution neural network with a customized target feature in the masked image modeling strategy:

- employing the vanilla convolution operator to construct the encoder-decoder architecture;
- implementing diverse lighting conditions by augmenting the input image with gamma transformation;
- dividing the corresponding illumination maps (obtained by the max intensity of RGB channels) into regular non-overlapping regions, randomly sampling a subset of regions, and masking the remaining ones while maintaining the whole structure of the map;
Enhancement of 

The solution of Eq. (10) is formulated as:

\[ L_{fp} = f_{MAEL}(L_{k+1}). \]  

(6)

Then, based on the prior features, an illumination map, \( L_{k+1} \), integrated with the customized illumination prior is formulated as:

\[ \hat{L}_{k+1} = L_{k+1} * \text{Sig}(\text{Conv}(L_{fp} \uparrow)) + \text{Sig}(\text{Conv}(L_{fp} \uparrow)), \]

(7)

where \( \uparrow \) denotes upsampling to the same spatial size as \( \hat{L}_{k+1} \). \( \text{Conv} \) indicates three convolution layers to generate the modulation parameter, and \( \text{Sig} \) is the Sigmoid function. The qualitative and quantitative evaluations of the effectiveness of the \( f_{MAEL} \) is presented in Fig. 8 and Table 4.

**R sub-problem:** Dropping the terms unrelated to \( R \) gives the following optimization problem:

\[
R_{k+1} = \arg\min_{R} \mathcal{L}\left( N_k, \hat{L}_{k+1}, R \right)
= \arg\min_{R} \omega \rho_3(R) + \frac{\mu_k}{2} \left\| h(R) - \frac{I - R \circ \hat{L}_{k+1} - N_k}{F} \right\|^2_F.
\]

(8)

Similarly, Eq. (8) is written as:

\[
R_{k+1} = \text{prox}_{\omega \rho_3}(R_k - \alpha_2 \nabla h(R_k)),
\]

(9)

where \( \text{prox}_{\omega \rho_3} \) is the proximal gradient operator corresponding to the prior \( \rho_3 \), implemented by two Conv layers followed by the ReLU activation, \( \alpha_2 \) indicates the updating step size, and \( \nabla h(R_k) = \mu_k \hat{L}_{k+1} \circ \left( \hat{L}_{k+1} \circ R_k + N_k - I \right) \).

**N sub-problem:** Collecting the terms related to \( N \) leads to the problem as follows:

\[
N_{k+1} = \arg\min_{N} \mathcal{L}\left( N, \hat{L}_{k+1}, R_{k+1} \right)
= \arg\min_{N} \left\| N \right\|_1 + \gamma \rho_1(N) + \frac{\mu_k}{2} \left\| I - R_{k+1} \circ \hat{L}_{k+1} - N \right\|^2_F.
\]

(10)

The solution of Eq. (10) is formulated as:

\[
N_{k+1} = \text{prox}_{\gamma \rho_1}(\text{Shrink}(I - R_{k+1} \circ \hat{L}_{k+1}, \frac{1}{\mu_k})),
\]

(11)

Fig. 5 presents the decomposed illumination, reflectance, and noise components of the low-/normal- light images, \( [L_l, R_l, N_l] \) and \( [L_n, R_n, N_n] \), respectively, where \( N_n \) is equal to zero. The second row illustrates the enhanced image, the enhanced illumination map, and the restored reflectance map, \( [L_{en}, L_{en}, R_{re}] \).

where \( \text{Shrink}(X, \eta) = \max\{\|X - \eta\|, 0\}\cdot\text{sign}(X) \) denotes the sign function, and \( \text{prox}_{\gamma \rho_1} \) indicates the proximal gradient operator corresponding to the prior \( \rho_1 \), achieved by two Conv layers followed by the ReLU activation.

Due to the absence of an optimal light level for images, a flexible illumination adjustment system is required. Following \[60\], an indicator with the value \( \varepsilon \) is combined with \( L_l \) as the input of the illumination adjustment network (\( \varepsilon = \text{mean}(L_n/L_l) \) for training and specified by users for inference). As shown in Fig. 2, the illumination adjustment network is implemented by a U-Net \[44\] with three scales of 12, 24, and 48 channels, respectively. Furthermore, the reflectance maps of the low-light images are degraded by color deviation and noise pollution \[60\], which are significantly associated with illumination and noise distribution. Thus, we integrate the illumination and noise maps, \( L_l \) and \( N_l \), with the degraded reflectance, \( R_l \), into the reflectance restoration network. Its structure is the same as the illumination adjustment network. Finally, the enhanced image, \( L_{en} \), will be obtained by element-wise multiplication between the enhanced illumination and the restored reflectance maps: \( L_{en} = L_{en} \circ R_{re} \). Fig. 5 presents the enhanced illumination, restored reflectance components, and the corresponding enhanced image.

3.1.2 Enhancement

After the Retinex decomposition, the low-light and normal-light images, \( I_l \) and \( I_n \), are decomposed into their illumination, reflectance, and noise terms, \( i.e., [L_l, R_l, N_l] \) and \( [L_n, R_n, N_n] \), where \( N_n \) is equal to zero.

It is combined with the input image \( L_l \) as the input of the illumination adjustment network. Following \[60\], an indicator with the value \( \varepsilon \) is combined with \( L_l \) as the input of the illumination adjustment network (\( \varepsilon = \text{mean}(L_n/L_l) \) for training and specified by users for inference). As shown in Fig. 2, the illumination adjustment network is implemented by a U-Net \[44\] with three scales of 12, 24, and 48 channels, respectively. Furthermore, the reflectance maps of the low-light images are degraded by color deviation and noise pollution \[60\], which are significantly associated with illumination and noise distribution. Thus, we integrate the illumination and noise maps, \( L_l \) and \( N_l \), with the degraded reflectance, \( R_l \), into the reflectance restoration network. Its structure is the same as the illumination adjustment network. Finally, the enhanced image, \( L_{en} \), will be obtained by element-wise multiplication between the enhanced illumination and the restored reflectance maps: \( L_{en} = L_{en} \circ R_{re} \). Fig. 5 presents the enhanced illumination, restored reflectance components, and the corresponding enhanced image.
3.2. Optimization Flow

Orthogonal to structure design, we introduce elaborately designed loss functions to enable the network for better optimization, thus producing pleasing images.

**Retinex decomposition:** Recovering illumination, reflectance, and noise terms simultaneously is an ill-posed problem. Fortunately, paired low-/normal-light images are available, and the reflectance of a certain scene should be shared across varying light conditions. Therefore, we maximize similarity between the decomposed reflectance pair, i.e., \( L^D = \| R_d - R_n \|_1 \). Furthermore, the illumination maps should be piece-wise smooth and mutually consistent. The illumination smoothness [60] is implemented by 

\[
L^I_{ts} = \frac{\| \nabla L_t \|_{1}}{\max(\| \nabla L_t \|)} + \frac{\| \nabla L_s \|_{1}}{\max(\| \nabla L_s \|)},
\]

where \( \nabla \) denotes the first order derivative operator in horizontal and vertical directions, \( | \cdot | \) means the absolute value operator, and \( \epsilon \) is set to 0.01 to avoid zero denominators. The mutual consistency [59] is formulated as 

\[
L^I_{mc} = \| M \circ \exp(-c \cdot M) \|_1
\]

with \( M = | \nabla L_t | + | \nabla L_s | \). It encourages the strong mutual edges to be well preserved and all weak edges to be effectively suppressed. In addition, the reconstruction error is constrained by 

\[
L^D_{re} = \| L - R_d \circ L_t - N_s \|_1 + \| L_n - R_d \circ L_n \|_1
\]

In total, the loss function of Retinex decomposition is as follows:

\[
L = L^D + L^E + L^N.
\]

**Enhancement:** The loss function in this step comprises illumination adjustment and reflectance restoration:

\[
L^E = \| R_{re} - R_n \|_1 + (1 - \text{SSIM}(R_{re}, R_n)) + \| L_{en} - L_n \|_1 + \| \nabla L_{en} - \nabla L_n \|_1,
\]

where \( \text{SSIM}(\cdot, \cdot) \) is the structure similarity measurement.

**Customized noise prior:** Low-light image enhancement focuses on correcting the luminance while eliminating the amplified noise after enhancement. Without noise removal operation, the inherent noise in low-light images will not be eliminated regardless of the adjusted lightness, i.e., being irrelevant to enhanced lightness. Histograms of Oriented Gradients [8] (HOG) describes the distribution of gradient orientations and magnitudes within a local sub-region while being innately irrelevant to enhanced lightness owing to the local normalization operation. Therefore, as shown in Fig. 6, we customize a noise prior with the MAE, where the target feature is the HOG feature of normal-light images:

- obtaining the HOG feature of the whole image;
- dividing the image into regular non-overlapping patches, randomly sampling a subset of patches, and masking the remaining ones;
- flattening the histograms of masked patches and concatenating them into a 1-D vector as the target feature;
- the encoder and decoder are implemented by vanilla vision transformers as in MAE.

After pre-training, the encoder of the noise prior, \( f_{MAE_n} \), possesses the powerful gradient representation capability. Thus, we redevelop it as a regularization term to suppress noise by imposing gradient representation consistency constraint between the enhanced and normal-light images, which is formulated as:

\[
L^N = \| f_{MAE_n}(I_{en}) - f_{MAE_n}(I_n) \|_1.
\]

The improvement of \( f_{MAE_n} \) in the noise elimination is illustrated in Fig. 8 and Table 4.

In total, the loss function of the proposed CUE for the end-to-end training paradigm is formulated as:

\[
L = L^D + L^E + L^N.
\]

4. Experiments

4.1. Experimental Settings

**Datasets.** To evaluate the performance of our proposed CUE, we conduct experiments on two commonly used benchmarks, including LOL [50] and Huawei [14]. LOL dataset contains 500 paired low-/normal images, and we split 485 for training and 15 for testing as the official selection. For the Huawei dataset, we randomly select 2200 images for training and the remaining 280 for testing.

**Implementations.** We implement our proposed method with PyTorch on a single NVIDIA GTX 3090 GPU. We employ the Adam optimizer with \( \beta_1 = 0.9, \beta_2 = 0.99 \) to update our model for 3000 steps with a batch of 16. The initial learning rate is set to \( 1 \times 10^{-4} \) and decreases with a factor of 0.5 every 50K iterations. The patch size is set to \( 64 \times 64 \). The coefficients in loss functions are consistent for training on both LOL and Huawei datasets.
Table 1. Quantitative comparison on the LOL and Huawei datasets in terms of PSNR↑, SSIM↑, and NIQE↓. The best and the second results are marked in bold and underlined, respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metrics</th>
<th>Method</th>
<th>SRIE</th>
<th>RetinexNet</th>
<th>MBLLEN</th>
<th>GLADNet</th>
<th>TBEFN</th>
<th>KinD</th>
<th>DRBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOL/Huawei</td>
<td>PSNR</td>
<td></td>
<td>12.28/13.40</td>
<td>16.77/16.65</td>
<td>17.56/16.63</td>
<td>19.72/17.76</td>
<td>17.35/16.88</td>
<td>20.87/16.48</td>
<td>18.65/18.46</td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td></td>
<td>0.596/0.477</td>
<td>0.425/0.485</td>
<td>0.729/0.526</td>
<td>0.680/0.521</td>
<td>0.781/0.575</td>
<td>0.802/0.540</td>
<td>0.801/0.635</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metrics</th>
<th>Method</th>
<th>URetinexNet</th>
<th>ZeroDCE++</th>
<th>EnGAN</th>
<th>RUAS</th>
<th>SCI</th>
<th>CUE (Ours)</th>
<th>CUE++ (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSIM</td>
<td></td>
<td>0.834/0.608</td>
<td>0.567/0.507</td>
<td>0.674/0.514</td>
<td>0.500/0.516</td>
<td>0.522/0.480</td>
<td>0.774/0.658</td>
<td>0.841/0.670</td>
</tr>
</tbody>
</table>

Figure 7. Visual comparison with state-of-the-art methods on the LOL and Huawei dataset. Please zoom in for details.

4.2. Comparison with State-of-the-Arts

We compare our CUE with 12 state-of-the-art approaches: SRIE [12], RetinexNet [50], MBLLEN [36], GLADNet [1], TBEFN [35], KinD [60], DRBN [54], URetinexNet [52], Zero-DCE++ [25], EnlightenGAN [61], RUAS [32], SCI [40] on the LOL and Huawei datasets.

**Quantitative results.** We employ the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and NIQE [43] as numerical evaluation metrics. Note that the URetinexNet [52] and KinD [60] are not trained end-to-end, where decomposition, unfolding, and enhancement stages are trained separately. For a fair comparison, we also implement our CUE by individually training the Retinex decomposition, reflectance restoration, and illumination adjustment stages and denote it as CUE++. To emphasize, CUE++ has the same model size, FLOPs, and runtime as CUE, and the only difference is the training strategy. As illustrated in Table 1, CUE/CUE++ achieves the best results on the two benchmarks, substantiating its effectiveness.

**Qualitative results.** Fig. 7 shows the subjective visual quality comparison on the challenging extremely low-light images from the Huawei dataset. TBEFN [35], GLADNet [1], and URetinexNet [52] generate the results with overexposure, while MBLLEN [36], SCI [40], and Zero-DCE++ [25] yield underexposed enhanced results. Undeniably, DRBN [54] and KinD [60] predict well-exposed images. However, the color distortion in their results is non-ignorable. RetinexNet [50], SCI [40], and TBEFN [35] produce the result with color deviation and artifacts. In contrast, our CUE effectively enhances lightness and reveals details while eliminating noise.

**Computational complexity.** We analyze the computational complexity of baselines and our CUE. We discard the traditional method (SRIE), and select SCI with the least parameters in three efficient LLIE approaches (Zero-DCE++, RUAS, and SCI) for comparison. Table 2 reports the model parameters, FLOPs, and runtime averaged on 50 images of 512×512. The running time is measured on a PC with an NVIDIA GTX 3070 GPU and an Intel i7-10700F CPU.
Table 2. Quantitative comparison in terms of Parameters, FLOPs, and Runtime.

<table>
<thead>
<tr>
<th>Method</th>
<th>RetinexNet</th>
<th>MBLENN</th>
<th>GLADNet</th>
<th>TBENF</th>
<th>KinD</th>
<th>DRBN</th>
<th>URetinexNet</th>
<th>EnGAN</th>
<th>SCI</th>
<th>CUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Params (M)</td>
<td>0.838</td>
<td>0.450</td>
<td>1.128</td>
<td>0.486</td>
<td>8.540</td>
<td>0.577</td>
<td>0.360</td>
<td>8.367</td>
<td>0.14</td>
<td>0.25</td>
</tr>
<tr>
<td>FLOPs (G)</td>
<td>148.54</td>
<td>21.37</td>
<td>275.32</td>
<td>26.33</td>
<td>36.57</td>
<td>42.41</td>
<td>233.09</td>
<td>72.61</td>
<td></td>
<td>157.32</td>
</tr>
<tr>
<td>Runtime (s)</td>
<td>0.023</td>
<td>0.159</td>
<td>0.024</td>
<td>0.035</td>
<td>0.068</td>
<td>0.140</td>
<td>0.115</td>
<td>0.011</td>
<td>0.104</td>
<td>0.104</td>
</tr>
</tbody>
</table>

Table 3. PSNR and SSIM scores of the proposed CUE with different numbers of stages on the LOL dataset.

<table>
<thead>
<tr>
<th>Stage (K)</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.06</td>
<td>0.739</td>
</tr>
<tr>
<td>2</td>
<td>21.48</td>
<td>0.767</td>
</tr>
<tr>
<td>3</td>
<td><strong>21.67</strong></td>
<td><strong>0.774</strong></td>
</tr>
<tr>
<td>4</td>
<td>21.65</td>
<td>0.771</td>
</tr>
<tr>
<td>5</td>
<td>21.54</td>
<td>0.766</td>
</tr>
</tbody>
</table>

Figure 8. Ablation studies of \( f_{MAEL} \) and \( f_{MAEN} \).

Compared with the baselines, our CUE performs best with a balanced model size and runtime.

4.3. Ablation Studies

We conduct ablation studies on the LOL dataset to further investigate the effectiveness of our proposed CUE under different configurations.

Effect of the number of unfolding stages: To explore the effect of the number of unfolding stages in the Retinex decomposition stage, we conduct experiments on proposed CUE with different numbers of stage \( K \). As illustrated in Table 3, the performance increases as the number of stages increases until reaching 3. The performance will decline when further increasing \( K \), which may be hampered by gradient propagation. We set \( K = 3 \) as the default stage number to balance the performance and computation cost. The optimal \( K \) and performance trend on the Huawei dataset is consistent with that of LOL (see supplementary material).

Effect of different components: We investigate the effectiveness of the proposed key components, i.e., \( f_{MAEL} \) and \( f_{MAEN} \). The quantitative and qualitative evaluations are presented in Table 4 and Fig. 8. \( f_{MAEL} \) denotes solving the L sub-problem with the encoder of the pre-trained illumination prior, and \( f_{MAEN} \) denotes employing the gradient representation prior as the regularization. As shown in Fig. 8, removing \( f_{MAEL} \) will cause the illumination map and the enhanced image to be over-smoothed. The enhanced image will suffer from severe noise pollution when discarding \( f_{MAEN} \). Combining them will achieve a notable performance improvement and a satisfying visual quality. Supplementation presents the detailed visual results.

Figure 9. Extension experiments on the effectiveness of \( f_{MAEN} \) for noise suppression. The numbers under each figure represent the PSNR/SSIM scores.

Table 4. PSNR and SSIM scores of the ablation studies for \( f_{MAEL} \) and \( f_{MAEN} \) on the LOL dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>( f_{MAEL} )</th>
<th>( f_{MAEN} )</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) DNCNN</td>
<td>35.71</td>
<td>0.911</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) DNCNN + ( f_{MAEN} )</td>
<td>38.10</td>
<td>0.952</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) MPRNet</td>
<td>41.31</td>
<td>0.974</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d) MPRNet + ( f_{MAEN} )</td>
<td>41.48</td>
<td>0.974</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4. Extension

We further verify the effectiveness of the proposed gradient representation consistency constraint in image denoising task. We conduct the experiments on the SIDD [2] dataset with two well-known baselines, DNCNN [58] and MPRNet [56]. Table 5 and Fig. 9 demonstrate that incorporating the gradient representation consistency constraint will improve the performance of the baselines.

Table 5. Extension experiments of the gradient representation consistency, \( f_{MAEN} \), on image denoising task. The performance is evaluated on the SIDD [2] dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNCNN [58]</td>
<td>32.43</td>
<td>0.841</td>
</tr>
<tr>
<td>DNCNN [58] + ( f_{MAEN} )</td>
<td><strong>32.68</strong></td>
<td><strong>0.857</strong></td>
</tr>
<tr>
<td>MPRNet [56]</td>
<td>39.25</td>
<td>0.956</td>
</tr>
<tr>
<td>MPRNet [56] + ( f_{MAEN} )</td>
<td><strong>39.32</strong></td>
<td><strong>0.957</strong></td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we have demonstrated the potential of customized learnable priors for enhancing low-light images and proposed a more transparent and interpretable deep unfolding paradigm. The customized illumination and noise priors are pre-trained using a masked image modeling strategy with specific target features. The learnable illumination prior is responsible for capturing the illumination proper-
ties of normal-light images, and we have embedded it into the unfolding architecture to improve its transparency in the structure flow. The noise prior aims to learn the gradient representation of normal-light images, and we have redeveloped it as a regularization term within the loss function to eliminate noise in the model output during the optimization flow. Extensive experiments on multiple low-light image datasets have shown the superiority of our deep unfolding paradigm.

Acknowledgments

This work was supported by the JKW Research Funds under Grant 20-163-14-LZ-001-004-01, and the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References


