Joint Low-light Enhancement and Super Resolution with Image Underexposure Level Guidance

Mingjie Xu¹
oxumingjie@buaa.edu.cn
Chaoqun Zhuang¹
zhuangchaoqun@buaa.edu.cn
Feifan Lv²
lvfeifan@buaa.edu.cn
Feng Lu¹*
lufeng@buaa.edu.cn

¹ State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University, Beijing, China
² Huawei Consumer BG Consumer Cloud Service Dept, Nanjing, China

Abstract

Obtaining high-quality images with high resolution in poor illumination environments using a limited spatial resolution image sensor poses a significant challenge. Low-light Enhancement (LLE) and Super-Resolution (SR) are crucial technologies for overcoming this challenge. However, current approaches usually generate normal-light high-resolution images with non-uniform brightness and loss of details from low-light low-resolution images, and suffer from significant performance degradation in cross-dataset settings. To alleviate these problems, we propose a novel solution for low-light image super-resolution. For non-uniform brightness problem, we propose a Relative Underexposure Level Estimation Module (RUL-EM) that estimates the relative underexposure levels of input images to adjust the image brightness to a uniform level and avoid artifacts. For detail loss and cross-dataset problems, we introduce the Multi-Scale Sampling (MSS) strategy for sampling multi-scale patches. MSS involves randomly cropping low-light and low-resolution patches of different sizes and positions and resizing them to a given patch size. Combining RUL-EM with MSS can improve the model performance in detail restoration and generalization. Additionally, we also incorporate channel attention to enable the Joint LLE & SR Network (JLSN) to adaptively adjust the influence of estimated relative underexposure levels. Our proposed method can be applied to various backbone architectures. Experimental results show that our proposed method achieves state-of-the-art performance on the joint LLE & SR task in both within-dataset and cross-dataset settings. Our proposed solution can convert low-resolution low-light images into high-resolution images with satisfactory brightness, vivid colors, and more details.

1 Introduction

In the real world, various image degradation problems significantly reduce image quality [28, 29] and create inconveniences for many kinds of applications, such as autonomous
driving [48], medical imaging [36], etc. Among these problems, two main issues are low-light (LL) and low-resolution (LR). Low-light problems are mainly caused by poor lighting conditions or limited exposure time when capturing images, while low-resolution problems are mainly caused by limited spatial resolution image sensors and image transmission over the Internet [34]. In a common scenario, images captured in low-light environments using regular cameras usually suffer from both low-light and low-resolution problems [2, 34].

In fact, obtaining high-quality images with high resolution in poor illumination environments using a limited spatial resolution image sensor is a highly challenging task. Both Low-Light Enhancement (LLE) [27, 45, 46] and Super-Resolution (SR) [23, 42] are crucial technologies for tackling this task. However, when inputting LL & LR images, the output normal-light (NL) & high-resolution (HR) images of current approaches are facing the following problems: (1) non-uniform brightness when inputting images of varying brightness, (2) loss of details and (3) significant performance degradation on unseen datasets, i.e., cross-dataset problem, which limits the generalization to the real world. These problems present a challenge in solving joint LLE & SR problem.

For the non-uniform brightness problem, it is crucial to incorporate the relative underexposure levels into the input to obtain uniform brightness from input images with varying brightness. Relative Underexposure Level (RUL) reflects the lightness of an image, ranging from \{0, 1, 2, 3, 4, 5, 6\}, with higher values denoting darker images and lower values indicating lighter images. Note that RUL is not a precise physical quantity. As such relative underexposure levels are difficult to obtain in the real world, we propose a Relative Underexposure Level Estimation Module (RUL-EM) to predict the accurate relative underexposure levels from LL & LR images, thereby enabling the model to adjust the image brightness to a uniform level from the input images of different brightness, and avoid artifacts. However, the RUL-EM alone cannot tackle the detail loss and the cross-dataset problems. We argue that these two problems are caused by the use of single-scale patches in the joint LLE & SR task, which is due to the fixed-size patch sampling strategy used for training (frequently used in low-level vision tasks). To overcome this, inspired by [50], we introduce the Multi-Scale Sampling (MSS) strategy [50] into our proposed method, which is used in other fields like image captioning [61] but rarely used in low-level vision. Different from [61], we randomly crop LL & LR patches of different sizes and positions, and resize them to a given patch size. Combining RUL-EM with MSS can improve detail restoration and generalization performance of the model, which is imperative for the joint LLE & SR task. Finally, we propose a Joint LLE & SR Network (JLSN) that uses channel attention [14] to enable the network to adaptively adjust the influence of the estimated relative underexposure levels, where various backbone architectures can be used. Based on these three components, we obtain a novel solution for low-light image super-resolution, which takes relative underexposure levels and multi-scale patches into consideration. Our proposed joint LLE & SR solution can alleviate the non-uniform brightness, the detail loss and the cross-dataset problems described above, as shown in Fig. 2, 3 and 5.

To summarize, our contributions are: 1) We propose a novel joint LLE & SR solution which can address the above-mentioned problems including non-uniform brightness, detail loss and cross-dataset problem. 2) We propose RUL-EM to accurately predict relative underexposure levels for adjustment of the image brightness to a uniform level and artifact avoidance, introduce MSS [61] to sample multi-scale patches of the scene and improve the ability of detail restoration and generalization with the help of RUL-EM, and propose JLSN that uses channel attention [14] to enable the network to adaptively adjust the influence of the estimated relative underexposure levels. 3) Both the quantitative and qualitative results
show that our proposed method can achieve state-of-the-art (SOTA) performance in the joint LLE & SR task in both within-dataset and cross-dataset settings, including all evaluation metrics, and produce high resolution images with satisfactory brightness, vivid colors, and more details.

2 Related Works

Image Super-Resolution. The Single Image Super-Resolution (SISR) task aims to increase the pixel density and enrich the details of an image. With the development of deep learning, SRCNN [6] first applied Convolutional Neural Network (CNN) to SISR. Since then, many works have made progress by stacking more convolutional layers and designing more complex network connections, like SRResNet [22], EDSR [24], RCAN [53], RDSR [17] and DCLS [26]. In order to solve the difficulty of signal-fidelity-oriented methods in reproducing image details, GAN [8] has a wide range of applications in the SISR task, including SRGAN [19], DBPN [12], ESRGAN [41], Real-ESRGAN [42] and LDL [22]. Recent SISR researches focus on Transformer [7, 25] architecture, such as SwinIR [23]. These methods perform well in recovering image details for normal-light low-resolution images. However, they cannot adjust image brightness and recover details well for low-light low-resolution images.

Low-light Image Enhancement (LLE). There exist various deep-learning-based LLE approaches, which are all able to achieve SOTA performance on popular benchmarks. These techniques typically rely on multi-level feature with multi-branch fusion [27], illumination cue [29], Retinex theory [18], pixel-wise and high-order curves [1, 21], multi-resolution features with spatial-wise and channel-wise attention [26], GAN [8], attention [30], self-calibrated illumination [31], Signal-to-Noise-Ratio prior [13], local-and-global components decoupling [3], local color distribution information [35] and over-and-under exposure consideration [35]. Although these approaches are able to tackle LLE problem well, their performance will degrade when applying them to joint LLE & SR tasks.

Joint LLE & SR. To the best of our knowledge, there are only a few methods that tackles joint LLE & SR task. Rasheed et al. proposed a joint LLE & SR network LSR using both the Lighten-Projection Unit and Darken-Projection Unit [34]. Aakerberg et al. also proposed a joint LLE & SR method named RELIEF [2]. Zhang et al. proposed a joint LLE & SR method for videos [47]. Guo et al. proposed a joint LLE & SR method for face images [10]. Moreover, RELLISUR dataset [1] is specifically designed for the image joint LLE & SR task. Such joint LLE & SR task is still open, which lacks research.

3 Methodology

In this section, we describe our proposed joint LLE & SR method. We begin by introducing the Relative Underexposure Level Estimation Module (RUL-EM) in Sec. 3.1, followed by the Multi-Scale Sampling (MSS) strategy in Sec. 3.2. Finally, we discuss the Joint LLE & SR Network (JLSN) in Sec. 3.3.

3.1 Relative Underexposure Level Estimation Module (RUL-EM)

Problem Analysis. Directly generating NL & HR images from LL & LR images can result in non-uniform brightness when inputting images with varying brightness. To tackle this problem, it is beneficial to use Relative Underexposure Levels (RUL) as input for the complex, degraded and coupled joint LLE and SR task. Relative underexposure levels are naturally and strongly correlated with brightness adjustment, which can help to enhance the
Obtaining relative underexposure levels for common RGB images distributed widely across the Internet is difficult. Thanks to the ground truth relative underexposure levels provided by the RELLISUR [1] dataset, we propose a Relative Underexposure Level Estimation Module (RUL-EM) to tackle such problem. Formally, this strategy can be expressed as \( \hat{y}_e = E(I_{LQ}) \), where \( \hat{y}_e \) is the estimated relative underexposure level and \( E \) is the RUL-EM.

**Implementation.** The RELLISUR dataset [1] provides 7 relative underexposure levels, which are only accessible in the training set. The goal of RUL-EM is to accurately predict such relative underexposure level of an input image. To achieve this, we design RUL-EM as a classification module, which is based on ResNet-50 [13] with Squeeze-and-Excitation module [14], as shown in Fig. 1. The model is trained using widely-used Cross Entropy Loss, and used to guide the subsequent MSS (Sec. 3.2) and JLSN (Sec. 3.3).

**Discussion.** Intuitively, RUL-EM pre-determines a certain relative underexposure level for a low-light low-resolution image. Based on this, the model learns a restricted problem, which is easier for avoiding artifacts.

### 3.2 Multi-Scale Sampling (MSS)

**Problem Analysis.** Although RUL-EM can help tackle the non-uniform brightness problem and avoid artifacts, experimental results have shown that RUL-EM alone can not tackle the detail loss and the cross-dataset problems (see Fig. 5). With the guidance of relative underexposure levels estimated by RUL-EM, there needs a strategy for detail restoration. Inspired by [50], we argue that using multi-scale patches with RUL-EM, instead of single-scale patches, can help alleviate the detail loss problem, further improving the cross-dataset performance.

Typically, current LLE or SR approaches sample fixed-size image patches for training [1, 21, 23, 45]. Such fixed-size patches are single-scale. For both single LLE or SR tasks, such single-scale patches contain enough information for color/illumination enhancement or detail restoration, respectively. However, single-scale patches are not enough for the joint LLE & SR task. If directly using single-scale patches for training the joint LLE & SR model, the model tends to generate over-smoothed and detail-lost images, as shown in Fig. 5. Although RUL-EM (Sec. 3.1) can help to avoid artifacts, using it alone cannot help the model for detail restoration. Moreover, using single-scale patches can cause the model to overfit as
they do not contain enough information for the joint LLE & SR task, resulting in significant performance degradation on unseen datasets. The supplementary material provides evidence that even large fixed-size patches are unable to address these issues. To tackle these issues, based on RUL-EM, multi-scale patches need to be applied into the training procedure. Thus, we introduce the Multi-Scale Sampling (MSS) strategy \cite{50} into our method for sampling multi-scale patches, which is used in other fields like image captioning \cite{50} but rarely used in low-level vision.

**Overview.** Different from \cite{50}, in brief, the MSS strategy randomly crops LL & LR patches with **different sizes and positions** and resizes them to a given patch size, as shown in Fig. 1. With the help of both RUL-EM and MSS, the joint LLE and SR model can easily obtain multi-scale patches, improving its ability of detail restoration and generalization on unseen datasets, which is imperative for the joint LLE & SR task.

**Procedure.** Given the full low-light low-resolution image $I_{LQ}$ with size $H \times W \times 3$, we randomly sample multiple patches $\mathcal{P} = \{P_1, P_2, \ldots\}$ with different sizes $p \times p \times 3$ where $p \sim \mathcal{U}(p_{\text{low}}, p_{\text{high}})$ and different top-left positions $x \sim \mathcal{U}(0, H - p)$ and $y \sim \mathcal{U}(0, W - p)$, where $\mathcal{U}$ is the Uniform distribution. Then we resize all the patches $P_k$ in $\mathcal{P}$ into the given size $s$, which outputs $\mathcal{P}^s = \{P_1^s, P_2^s, \ldots\}$ with size $s \times s \times 3$.

**Discussion.** MSS addresses the cross-dataset generalization problem by sampling patches with a wide range of scales, making the model learn various scales that may appear in unseen datasets during training. Without MSS, the model is unable to effectively learn the diverse scales present in unseen datasets since the training dataset only covers a limited range of scales across different input images.

### 3.3 Joint LLE & SR Network (JLSN)

**Overview.** In the joint LLE & SR procedure, the relative underexposure levels estimated by RUL-EM (Sec. 3.1) and the multi-scale patches sampled by MSS strategy (Sec. 3.2) are helpful for restoring NL & HR images $\hat{I}_{HQ}$ from LL & LR images $I_{LQ}$. Here we propose a Joint LLE & SR Network (JLSN) to restore NL & HR images with the guidance of such information, which is shown in Fig. 1. The JLSN consists of a Generator $G$ and Discriminator $D$ following the classic GAN \cite{8} structure, and can use various backbone architectures.

**Generator $G$.** Our proposed method is in the form of add-on (plug-and-play), which constitutes one of its key advantages. We employ commonly-used LLE architecture MIRNet \cite{46} (with upsampling module \cite{37}, a representative LLE backbone), SR architectures, including RRDB \cite{41, 42} (a representative CNN-based backbone) and SwinIR \cite{23} (a representative Transformer-based backbone), and a cascaded LLE & SR architecture MIRNet+RRDB \cite{46, 41, 42} as $G$. To fully exploit the connection between the relative underexposure levels and the LL & LR image $I_{LQ}$, we incorporate the Channel Attention (CA) structure \cite{14} into RRDB, SwinIR and MIRNet+RRDB Module (MIRNet+Upsample originally contains CA), allowing the JLSN to adjust the influence of the estimated relative underexposure levels adaptively. We refer to the modified generator models as CA-RRDB, CA-SwinIR and MIRNet+CA-RRDB, respectively. Implementation details can be found in the supplementary material. Further experiments show that plugging our proposed method into various backbone can all achieve SOTA results and does not increase the inference time, as shown in Sec. 4.2 and the supplementary material, respectively.

**Discriminator $D$.** Here we use the U-Net Discriminator with spectral normalization following \cite{42} to increase the adaptability of the network to real-world scenarios.

**RUL-EM and MSS Information Combination.** To make use of the information obtained by RUL-EM and MSS, we replicate the estimated relative underexposure level $y_e$ into the
relative underexposure map $\mathbf{M}$ with size $s \times s \times 1$. We then concatenate $\mathbf{M}$ and the multi-scale patch $\mathbf{P}_k$ to obtain the input $\mathbf{I}_k = \mathbf{P}_k \oplus \mathbf{M}, \mathbf{P}_k \in \mathcal{P}^s$ to $G$, where $\oplus$ indicates the concatenate operation. Note that this procedure is only processed in the training phase. To simplify the implementation, we randomly select one $p$ and one $(x, y)$ for each image in one batch. In the testing phase, we directly feed the full $\mathbf{I}_{LQ}$ concatenated with the estimated relative underexposure map (replicated using the estimated relative underexposure level $\hat{y}_e$) into the JLSN to generate the normal-light high-resolution image.

**Loss Function.** We use $L_1$ Pixel Loss $\mathcal{L}_{pix}$, Perceptual Loss $\mathcal{L}_{per}$ \cite{7, 8} and Adversarial Loss $\mathcal{L}_{adv}$ \cite{8} for training $G$, and the Adversarial Loss $\mathcal{L}_D$ \cite{8} for training $D$, which is commonly used in both LLE methods and SR methods. The total loss function of the generator $G$ is $\mathcal{L}_G = \mathcal{L}_{pix} + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv}$, where $\lambda_{per}$ and $\lambda_{adv}$ are the coefficients to balance different terms of $\mathcal{L}_G$, and we empirically set them as $\{1.0, 0.005\}$ respectively.

## 4 Experiments

### 4.1 Setup

**Dataset.** We use the RELLISUR ($\mathcal{D}_R$) dataset \cite{2} for within-dataset evaluation, which provides real-world LL & LR images of 7 different relative underexposure levels and NL & HR images with $\times 1, \times 2$ or $\times 4$ magnifications, which can be used to model real-world physical down-sampling and noise. There are 3610, 215, and 425 images for training, validation, and testing, respectively. Moreover, to evaluate the cross-dataset performance, we also use LOL ($\mathcal{D}_L$) \cite{44}, LSRW ($\mathcal{D}_S$) \cite{11} and DIV2K ($\mathcal{D}_D$) \cite{3, 38} test sets, which provide 15, 50 and 100 images, respectively. This forms 3 cross-dataset tasks: $\mathcal{D}_R \rightarrow \mathcal{D}_L$, $\mathcal{D}_R \rightarrow \mathcal{D}_S$ and $\mathcal{D}_R \rightarrow \mathcal{D}_D$, respectively. Since LOL \cite{44} and LSRW \cite{11} are designed for LLE task, they only contain NL & LR ground truth images. Since DIV2K \cite{3, 38} is designed for SR task, we perform Gamma correction with $\gamma \sim \mathcal{U}(2, 3.5)$ on the input LR images following \cite{27}.

**Comparison Methods.** Our task is joint LLE and SR. So it is necessary to compare our proposed method with SOTA LLE and SR methods. We conduct a comparison with SOTA and latest methods on single LLE or SR tasks on RELLISUR dataset. Results are shown in the supplementary material. Based on these results, we select the LLE methods SNR-Aware \cite{45}, MBLLEN \cite{27} and MIRNet \cite{46}, as well as the SR methods SRResNet \cite{20}, EDSR \cite{24}, Real-ESRGAN \cite{42}, SwinIR \cite{23} and LDL \cite{22} which have the best performances, for further experiments and comparison. Moreover, we compare our proposed method with SOTA joint LLE & SR methods LSR \cite{34} and RELIEF \cite{2}. We compare our proposed method with 6 strategy designs: (I) Apply LLE methods to obtain NL & LR images from LL & LR images, followed by SR methods to obtain the desired NL & HR images, (II) Reversed Sequential Process of Type I, (III) Cascaded LLE Network and SR Network, (IV) LLE Network + Upsampling Module \cite{37}, (V) SR Network, and (VI) Joint LLE & SR Network. Implementation details will be discussed in the supplementary material.

Our proposed method is in the form of add-on (plug-and-play). To demonstrate the SOTA performance of our proposed method with different backbones, we use 4 versions of our proposed method with different backbones for comparison, including CA-RRDB, CA-SwinIR, MIRNet+CA-RRDB and MIRNet+Upsample, as stated in Sec. 3.3.

**Evaluation Metrics.** We use Peak Signal to Noise Ratio (PSNR) \cite{15} and Structural Similarity (SSIM) \cite{43} to evaluate the similarity between a HR image and its generated SR counterpart (higher is better), while we use Learned Perceptual Image Patch Similarity (LPIPS) \cite{49} and Natural Image Quality Evaluator (NIQE) \cite{32} to evaluate human perceptual quality of
Table 1: Quantitative results of SOTA methods. ↑ indicates lower is better, ↓ indicates lower is better. Red, blue and green denotes the first, second and third best results, respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR-Aware</td>
<td>SNR-Aware</td>
<td>20.19</td>
<td>0.75</td>
<td>3.2</td>
<td>7.46</td>
<td>19.49</td>
<td>0.75</td>
<td>3.2</td>
<td>7.46</td>
<td>6.79</td>
<td>0.74</td>
<td>3.2</td>
<td>6.79</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
</tr>
<tr>
<td>MIRNet [28]</td>
<td>MIRNet [28]</td>
<td>21.46</td>
<td>0.75</td>
<td>4.7</td>
<td>8.79</td>
<td>20.49</td>
<td>0.76</td>
<td>3.7</td>
<td>10.59</td>
<td>8.7</td>
<td>0.78</td>
<td>3.7</td>
<td>10.59</td>
<td>8.7</td>
<td>8.7</td>
<td>8.7</td>
<td>8.7</td>
<td>8.7</td>
<td>8.7</td>
<td>8.7</td>
<td>8.7</td>
</tr>
<tr>
<td>MBLLEN [27]</td>
<td>MBLLEN [27]</td>
<td>18.88</td>
<td>0.74</td>
<td>4.9</td>
<td>8.27</td>
<td>18.23</td>
<td>0.74</td>
<td>5.7</td>
<td>10.43</td>
<td>8.5</td>
<td>0.83</td>
<td>5.7</td>
<td>10.43</td>
<td>8.5</td>
<td>8.5</td>
<td>8.5</td>
<td>8.5</td>
<td>8.5</td>
<td>8.5</td>
<td>8.5</td>
<td>8.5</td>
</tr>
<tr>
<td>SNR-Aware</td>
<td>SNR-Aware</td>
<td>20.33</td>
<td>0.75</td>
<td>3.3</td>
<td>7.67</td>
<td>19.15</td>
<td>0.75</td>
<td>4.8</td>
<td>9.06</td>
<td>7.1</td>
<td>3.2</td>
<td>4.8</td>
<td>9.06</td>
<td>7.1</td>
<td>7.1</td>
<td>7.1</td>
<td>7.1</td>
<td>7.1</td>
<td>7.1</td>
<td>7.1</td>
<td>7.1</td>
</tr>
<tr>
<td>MIRNet [27]</td>
<td>MIRNet [27]</td>
<td>21.45</td>
<td>0.75</td>
<td>4.6</td>
<td>9.86</td>
<td>20.08</td>
<td>0.75</td>
<td>5.9</td>
<td>9.88</td>
<td>8.4</td>
<td>3.3</td>
<td>5.9</td>
<td>9.88</td>
<td>8.4</td>
<td>8.4</td>
<td>8.4</td>
<td>8.4</td>
<td>8.4</td>
<td>8.4</td>
<td>8.4</td>
<td>8.4</td>
</tr>
<tr>
<td>SNR-Aware</td>
<td>SNR-Aware</td>
<td>18.88</td>
<td>0.76</td>
<td>4.9</td>
<td>8.56</td>
<td>17.79</td>
<td>0.74</td>
<td>5.8</td>
<td>9.82</td>
<td>8.9</td>
<td>3.9</td>
<td>5.8</td>
<td>9.82</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
<td>8.9</td>
</tr>
<tr>
<td>MIRNet [28]</td>
<td>MIRNet [28]</td>
<td>20.02</td>
<td>0.75</td>
<td>3.0</td>
<td>6.91</td>
<td>19.13</td>
<td>0.75</td>
<td>4.1</td>
<td>7.40</td>
<td>5.1</td>
<td>3.2</td>
<td>4.1</td>
<td>7.40</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
</tr>
<tr>
<td>MIRNet [28]</td>
<td>MIRNet [28]</td>
<td>21.50</td>
<td>0.74</td>
<td>4.1</td>
<td>7.65</td>
<td>20.48</td>
<td>0.75</td>
<td>5.1</td>
<td>8.85</td>
<td>6.6</td>
<td>4.9</td>
<td>5.1</td>
<td>8.85</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
</tr>
<tr>
<td>SNR-Aware</td>
<td>SNR-Aware</td>
<td>18.87</td>
<td>0.74</td>
<td>4.2</td>
<td>9.70</td>
<td>18.12</td>
<td>0.73</td>
<td>5.1</td>
<td>8.67</td>
<td>5.9</td>
<td>4.0</td>
<td>5.1</td>
<td>8.67</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
</tr>
<tr>
<td>SNR-Aware</td>
<td>SNR-Aware</td>
<td>20.31</td>
<td>0.75</td>
<td>3.3</td>
<td>7.46</td>
<td>19.46</td>
<td>0.75</td>
<td>4.8</td>
<td>9.32</td>
<td>7.4</td>
<td>4.1</td>
<td>4.8</td>
<td>9.32</td>
<td>7.4</td>
<td>7.4</td>
<td>7.4</td>
<td>7.4</td>
<td>7.4</td>
<td>7.4</td>
<td>7.4</td>
<td>7.4</td>
</tr>
<tr>
<td>MIRNet [28]</td>
<td>MIRNet [28]</td>
<td>21.44</td>
<td>0.75</td>
<td>4.6</td>
<td>9.84</td>
<td>20.48</td>
<td>0.76</td>
<td>5.9</td>
<td>10.37</td>
<td>8.8</td>
<td>3.8</td>
<td>5.9</td>
<td>10.37</td>
<td>8.8</td>
<td>8.8</td>
<td>8.8</td>
<td>8.8</td>
<td>8.8</td>
<td>8.8</td>
<td>8.8</td>
<td>8.8</td>
</tr>
<tr>
<td>MIRNet [28]</td>
<td>MIRNet [28]</td>
<td>18.83</td>
<td>0.75</td>
<td>4.6</td>
<td>9.11</td>
<td>18.64</td>
<td>0.75</td>
<td>5.1</td>
<td>9.74</td>
<td>5.0</td>
<td>4.0</td>
<td>5.1</td>
<td>9.74</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>SNR-Aware</td>
<td>SNR-Aware</td>
<td>19.99</td>
<td>0.75</td>
<td>3.0</td>
<td>6.78</td>
<td>19.35</td>
<td>0.76</td>
<td>4.1</td>
<td>8.32</td>
<td>4.9</td>
<td>3.4</td>
<td>4.1</td>
<td>8.32</td>
<td>4.9</td>
<td>4.9</td>
<td>4.9</td>
<td>4.9</td>
<td>4.9</td>
<td>4.9</td>
<td>4.9</td>
<td>4.9</td>
</tr>
<tr>
<td>SRResNet</td>
<td>SRResNet [20]</td>
<td>21.43</td>
<td>0.74</td>
<td>4.1</td>
<td>7.73</td>
<td>20.40</td>
<td>0.75</td>
<td>5.1</td>
<td>8.84</td>
<td>6.4</td>
<td>4.3</td>
<td>5.1</td>
<td>8.84</td>
<td>6.4</td>
<td>6.4</td>
<td>6.4</td>
<td>6.4</td>
<td>6.4</td>
<td>6.4</td>
<td>6.4</td>
<td>6.4</td>
</tr>
<tr>
<td>MIRNet [28]</td>
<td>MIRNet [28]</td>
<td>18.80</td>
<td>0.74</td>
<td>4.2</td>
<td>7.73</td>
<td>18.11</td>
<td>0.74</td>
<td>5.1</td>
<td>8.71</td>
<td>6.6</td>
<td>5.8</td>
<td>5.1</td>
<td>8.71</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
<td>6.6</td>
</tr>
</tbody>
</table>

We use PyTorch to conduct all the experiments on NVIDIA GPUs. All proposed models are trained for 200,000 iterations with a batch size of 24. For both G and D, we use the Adam optimizer with $lr = 10^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and set the weight decay as 0. The LR patch size $s$ is set to 64 for all the experiments. We set the dimension of the feature of both $G$ and $D$ as 64. For CA-RRDB, we use 7 blocks, and the channel growth of the dense block is set to 32. For other architectures, we use their original settings. We use the same setting of Perceptual Loss as Real-ESRGAN [23]. For MSS, we set $p_{low}$ as 64 and $p_{high}$ as 512. For RUL-EM, the model is pretrained before training other modules. The input size is 625 as it is not cropped, and the model is trained for 200 epochs with the SGD optimizer with $lr = 10^{-4}$, $\gamma = 0.7$, and we set the weight decay as $5 \times 10^{-4}$.  

4.2 Comparison with SOTA Methods

To show the superior performance of our proposed method on joint LLE & SR task, we conduct comparison experiments with SOTA methods on RELLISUR dataset [30]. Quantitative and qualitative results are shown in Tab. 1 and Fig. 2, respectively. In Tab. 1, our proposed joint LLE & SR methods can outperform all types of SOTA methods on all evaluation met-
Figure 2: Qualitative results on RELLISUR dataset (x4). Zoom for best view. Our proposed methods tend to generate vivid colors and more details.

Figure 4(a): The confusion matrix for RUL-EM on RELLISUR test split. The model achieves 77.8% accuracy.
always accurately predict the relative underexposure levels, it can predict them similar to ground truth for most of the time. So RUL-EM can be used in subsequent procedures.

**Ablation Study.** To further show the efficacy of our proposed method and the necessity of all its components, we conduct an ablation study using different combinations of components. All the experiments are conducted using the CA-RRDB backbone, which is a common and simple-structured CNN in LLE or SR and can represent prevalent cases. Quantitative and qualitative results are shown in Fig. 4(b) and Fig. 5 respectively. With the help of the RUL-EM, the quality of output $\hat{I}_{HQ}$ significantly increases, as indicated by the increased PSNR and SSIM values, and the joint LLE & SR network tends to adjust the input low-light images to uniform brightness. Additionally, RUL-EM can help the network to avoid some artifacts. However, using RUL-EM alone cannot tackle the detail loss problem. Based on RUL-EM, if MSS is combined into the network, an image with uniform brightness with more details and fewer artifacts can be obtained than RUL-EM alone. Moreover, using RUL-EM with Channel Attention (CA) can further improve the PSNR, SSIM and LPIPS values, and make the network product more vivid colors and fewer artifacts. This is mainly because CA can

Figure 3: Qualitative results on LOL dataset (x4). Zoom for best view. Our proposed methods tend to generate **vivid colors, more details and fewer artifacts** on unseen datasets.
Table 1: Performance of different strategies on joint LLE & SR task. ↑ indicates larger is better, ↓ indicates lower is better.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
<th>NIQE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>18.59</td>
<td>0.76</td>
<td>0.40</td>
<td>8.32</td>
</tr>
<tr>
<td>RUL-EM</td>
<td>20.97</td>
<td>0.78</td>
<td>0.40</td>
<td>8.18</td>
</tr>
<tr>
<td>RUL-EM + MSS</td>
<td>20.76</td>
<td>0.76</td>
<td>0.42</td>
<td>5.85</td>
</tr>
<tr>
<td>RUL-EM + CA</td>
<td>21.66</td>
<td>0.79</td>
<td>0.39</td>
<td>8.62</td>
</tr>
<tr>
<td>RUL-EM + MSS + CA</td>
<td>21.52</td>
<td>0.77</td>
<td>0.39</td>
<td>6.77</td>
</tr>
</tbody>
</table>

Figure 4: (a) Confusion matrix of RUL-EM on the test split of RELLISUR dataset, percentage is shown. (b) Ablation study on different strategies on joint LLE & ×4 SR task. ↑ indicates larger is better, ↓ indicates lower is better.

Figure 5: Qualitative results of ablation study (x4). Zoom for best view.

help the network better handle the relationship between the relative underexposure levels and the input. Finally, using both RUL-EM and MSS with CA can obtain the most details and the colors closest to the ground truth with the most uniform brightness. Note that MSS may cause a slight decrease in PSNR and SSIM values. This is because details and PSNR/SSIM values are a pair of trade-off as noted in [19]. Higher PSNR/SSIM values tend to be related to over-smoothing problems, while more details may cause lower PSNR/SSIM values [19]. It can be concluded from the ablation study that all the components are necessary for the joint LLE & SR task.

5 Conclusion

In this paper, we propose a novel solution for the joint LLE and SR task. We propose a Relative Underexposure Level Estimation Module (RUL-EM) to accurately estimate relative underexposure levels for adjusting the image brightness to a uniform level and artifact avoidance. Furthermore, we introduce the efficient Multi-Scale Sampling (MSS) strategy [50] that enables the network to sample multi-scale patches of one scene. Cooperating RUL-EM and MSS can improve the detail restoration and generalization performance. Lastly, we design a Joint LLE & SR Network (JLSR), incorporating Channel Attention (CA) into the various architectures to adaptively adjust the influence of the estimated relative underexposure levels. Experimental results show that our proposed method achieves the SOTA performance on RELLISUR, LOL, and LSRW datasets with the highest quality, vivid colors, more details, and fewer artifacts in both within-dataset and cross-dataset settings. This paper can help for more robust applications of computer vision techniques in extreme environments.

References


