Low-Light Image Enhancement via Structure Modeling and Guidance

Xiaogang Xu\textsuperscript{1} \quad Ruixing Wang\textsuperscript{2} \quad Jiangbo Lu\textsuperscript{3*}
\textsuperscript{1} Zhejiang Lab \quad \textsuperscript{2} Honor Device Co., Ltd. \quad \textsuperscript{3} SmartMore Corporation
xgxu@zhejianglab.com, ruixingw@hustunique.com, jiangbo@smartmore.com

Abstract

This paper proposes a new framework for low-light image enhancement by simultaneously conducting the appearance as well as structure modeling. It employs the structural feature to guide the appearance enhancement, leading to sharp and realistic results. The structure modeling in our framework is implemented as the edge detection in low-light images. It is achieved with a modified generative model via designing a structure-aware feature extractor and generator. The detected edge maps can accurately emphasize the essential structural information, and the edge prediction is robust towards the noises in dark areas. Moreover, to improve the appearance modeling, which is implemented with a simple U-Net, a novel structure-guided enhancement module is proposed with structure-guided feature synthesis layers. The appearance modeling, edge detector, and enhancement module can be trained end-to-end. The experiments are conducted on representative datasets (sRGB and RAW domains), showing that our model consistently achieves SOTA performance on all datasets with the same architecture. The code is available at https://github.com/xiaogang00/SMG-LLIE.

1. Introduction

The low-light enhancement aims to recover normal-light and noise-free images from dark and noisy pictures, which is a long-standing and significant computer vision topic. It has broad application fields, including low-light imaging \cite{11, 27, 50}, and also benefits many downstream vision tasks, e.g., nighttime detection \cite{28, 48, 64}. Some methods have been proposed to tackle the low-light enhancement problem. They design networks that learn to manipulate color, tone, and contrast \cite{7, 10, 45, 63}, and some recent works also account for noise in images \cite{24, 55}. Most of these works optimize the appearance distance between the output and the ground truth. However, they ignore the explicit modeling of structural details in dark areas and thus resulting in blurry outcomes and low SSIM \cite{51} values, as shown in Fig. 2. Some works \cite{34, 74} have noticed the effect of employing structural information, e.g., edge, to promote the enhancement. Edge guides the enhancement by distinguishing between different parts in the dark regions. Moreover, adding sensible edge priors into dark regions reduces the ill-posed degree in optimizing the appearance reconstruction. These frameworks \cite{34, 74} perform the structure modeling with encoder-decoder-based networks and a regression loss. However, the corresponding structure modeling results are not satisfying due to the uncertainty in the dark areas caused by severely poor visibility and noise. Furthermore, the strategy of using the extracted structural information needs to be improved from the existing straightforward concatenation approach \cite{34, 74}.

In this paper, we propose to utilize a generative model $S$ trained with a GAN loss to perform the structure modeling with the form of edges. Then, we design a new mechanism $\mathcal{E}$ to facilitate the initial low-light appearance enhancement (the module is denoted as $A$) with structure-guided feature synthesis. With effective structure modeling and guidance, our framework can output sharp and realistic results with satisfactory reconstruction quality as shown in Fig. 2. Compared with previous structure modeling networks,
the proposed generative model $S$ has two significant modifications. First, we notice the impact of providing structure-aware descriptors into both the encoder and decoder of $S$, disentangling the appearance representation and underlining the structural information. Thus, we design a Structure-Aware Feature Extractor (SAFE) as the encoder part, which extracts structure-aware features from the dark image and its gradients via spatially-varying operations (achieved with adaptive long-range and short-range computations). The extracted structure-aware tensors are then fed into the decoder part to generate the desired structure maps. Moreover, different from current approaches, which employ the structure maps of normal-light images to conduct the regression learning, we find the nice property of using a GAN loss. The GAN loss can reduce the artifacts in the generated structure maps that are caused by the noise and invisibility, highlighting the essential structure required for enhancement. The backbone of $S$ is implemented as a modified StyleGAN.

To boost the appearance by leveraging the obtained structure maps, we design a Structure-Guided Enhancement Module (SGEM) as $\mathcal{E}$. The main target of SGEM is to learn the residual, which can improve the initial appearance modeling results. In SGEM, spatially-adaptive kernels and normalization parameters are generated according to the structure maps. Then, the features in each layer of the SGEM’s decoder will be processed with spatially-adaptive convolutions and normalization. Although the overall architecture of SGEM takes the form of a simple U-Net [38], it can effectively enhance the original appearance.

$S$, $\mathcal{A}$, and $\mathcal{E}$ can be trained end-to-end simultaneously. Extensive experiments are conducted on representative benchmarks. Experimental results show that our framework achieves SOTA performance on both PSNR and SSIM metrics with the same architecture on all datasets, as shown in Fig. 1. In summary, our work’s contribution is four-fold.

1. We propose a new framework for low-light enhancement by conducting structure modeling and guidance simultaneously to boost the appearance enhancement.

2. We design a novel structure modeling method, where structure-aware features are formulated and trained with a GAN loss.

3. A novel structure-guided enhancement approach is proposed for appearance improvement guided by the restored structure maps.

4. Extensive experiments are conducted on different datasets in both sRGB and RAW domains, showing the effectiveness and generalization of our framework.

### 2. Related Work

**Low-light enhancement with learning.** The current low-light image enhancement methods with deep learning mainly optimize the appearance reconstruction error between the output and the ground truth [1, 9, 13, 17, 24, 26, 46, 55, 57, 58, 61, 62, 67, 68, 70, 71]. However, the enhanced appearance tends to be blurry in the dark regions even with SOTA approaches [56]. To this, some works utilize the structural information [16, 74] to enhance appearance. To obtain the structure maps, some of them use offline-computed edges/gradadients [11, 16, 36, 41] or offline-trained networks [23] which are not adaptively optimized with the low-light data and result in artifacts inevitably. Although several methods [34, 74] have been proposed to train an edge detector and an image enhancement network simultaneously with both regression losses, their improvements are still limited. Moreover, existing strategies mainly concatenate the extracted structure maps with the images to facilitate the enhancement [16, 34, 74], which, however, cannot set the guidance in all layers of the enhancement network. Different from current works, we propose a new generative model to robustly extract the edge map from a dark input image with the GAN loss, providing the critical edge information for enhancement in every feature layer.

**Generative model for restoration.** The recent development on the generative model enables a wide range of networks, e.g., StyleGAN [15], to achieve significant restoration effects in some tasks, e.g., face restoration [12, 59, 73]. However, their performances are highly dependent on the pre-trained model within several categories. In this paper, we demonstrate that generative models with the GAN loss can be utilized in a novel way to synthesize structural information for the restoration task of low-light image enhancement. Such an approach does not rely on the pre-trained model, and such an idea can actually be extended to various restoration tasks, opening up a new research direction.
3. Method

The overview of our framework can be viewed in Fig. 3. The proposed framework conducts explicit modeling for both appearance \( A \) as well as structure \( S \), and they will be described in Sec. 3.1 and Sec. 3.2, respectively. The structure maps are employed to guide the enhancement of appearance modeling with the module of \( E \), which will be introduced in Sec. 3.3. The overall framework can be trained end-to-end as elaborated in Sec. 3.4.

3.1. Appearance Modeling

In our framework, given an input low-light image \( I \), we adopt a common U-Net \cite{3} as the network for appearance modeling, as shown in Fig. 3 (a), where \( I_a = A(I) \). The experimental results show that our approach combined with the structural modeling can achieve SOTA performance, with such an ordinary network for appearance prediction.

3.2. Structure Modeling

In this paper, we adopt edge information for structure modeling, which is a general representation applicable to various scenes. We find that the effective generative model, e.g., the StyleGAN \cite{15}, can be modified to be a strong structure estimator \( S \) for low-light images as shown in Fig. 3 (b). Equipped with both the structure regression loss and the GAN loss, \( S \) can accurately predict the effective structure maps that are beneficial for appearance enhancement. In this paper, we use the backbone of StyleGAN to formulate \( S \). Given the input image \( I \), we can obtain its structure map \( S \) as

\[
I_s = S(I) = G(F(I)),
\]

where \( F \) is the encoder part, as “Structure-Aware Feature Extractor (SAFE)” that contains \( N \) layers and two mapping functions, and \( G \) is decoder part, as “Structure-Aware Generator (SAG)”. \( F \) and \( G \) enable our framework to perform better structure modeling than previous generative models and edge detectors.

3.2.1 Structure-Aware Feature Extractor (SAFE)

Although the traditional encoder-based generative model can be employed as the structure extractor, its performance can be improved by feeding structural features into both the encoder and the generator. Previous encoder-based generative models extract information from only the content feature, which is suitable for the content generation (e.g., sRGB images) while insufficient for structure modeling. Moreover, the conventional feature extractor of the generative model mainly utilizes the short-range operation \cite{37, 42, 47}, e.g., CNN. However, as indicated in \cite{56}, formulating useful representations in dark areas with low Signal-to-Noise Ratio (SNR) needs long-range operations.

Thus, we design the SAFE in \( S \) as shown in Fig. 3 (b), whose main target is to extract the reliable structural features that can be utilized next in the SAG. Different from conventional generative models’ encoders, SAFE obtains the required information from both content and gradient maps which are helpful for edge modeling. Spatially-varying operations are formulated in SAFE for feature extraction.

The formulation of gradient maps. As shown in Fig. 3 (b), the input image \( I \) is extracted by a multi-layer encoder, \( F_1, ..., F_N \), with down-sampling in each layer. Suppose
the input feature of each layer is $f_i$, $i \in [1, N]$, we obtain the gradient map of $f_i$ in the directions of $x$- and $y$-axis, as $\{g_{+x}(f_i), g_{-x}(f_i), g_{+y}(f_i), g_{-y}(f_i), g_{+x+y}(f_i), g_{+x-y}(f_i), g_{-x+y}(f_i), g_{-x-y}(f_i)\}$, where $g$ is the function to compute the first order gradient. These gradient maps emphasize the edge regions, benefiting the formulation of the structural features.

**Spatially-varying feature extraction for structure.** We design the spatially-varying operations in the encoder to effectively extract structural features. For $f_i$ and $g(f_i)$, we set the corresponding Long-Range Encoder (LRE) and Short-Range Encoder (SRE) modules. The transformer blocks are proven to have extraordinary ability in the long-range modeling [4, 25, 52, 53, 65, 66], and the CNN blocks are capable of extracting short-range characteristics. Thus, LRE and SRE are implemented as the transformer and CNN blocks, respectively. To build an efficient transformer module, we adopt the transformer block with the window-based attention mechanism [25] and local-enhanced forward module [52]. Suppose LRE and SRE modules are denoted as $F_i$ and $F_s$, the feature extraction process can be written as

$$
    l_i = F_i^l(f_i), \quad s_i = F_s^l(f_i), \quad \nabla l_i = \nabla F_i^l(\nabla g(f_i)), \quad \nabla s_i = \nabla F_s^l(\nabla g(f_i)),
$$

(2)

where $\nabla \in \{+, −, +x, −x, +y, −y, +x+y, +x-y, −x+y, −x-y\}$, $l_i$ is the long-range feature and $s_i$ is the short-range feature, respectively. We adopt another Long-Short-Range Fusion (LSR-F) module to merge the outputs from long-range and short-range operations, whose architecture is the multilayer perceptron. Suppose LSR-F is denoted as $F_{i,l}^l$, then the spatially-varying extracted features for $f$ and $g$ can be formulated as

$$
    h_i = F_{i,l}^l(l_i, s_i), \quad \nabla h_i = \nabla F_{i,l}^l(\nabla l_i, \nabla s_i).
$$

(3)

Moreover, these spatially-varying extracted features for different directions can be adaptively merged with the Gradient Fusion (Grad-F) module, as

$$
    f_{i+1} = F_{i,l}^{l}(h_i, +x h_i, −x h_i, −y h_i, +y h_i, +x y h_i, +x y h_i, −x y h_i, −x y h_i),
$$

(4)

where $F_{i,l}^{l}$ denotes the Grad-F module.

3.2.2 Structure-Aware StyleGAN Generator (SAG)

After obtaining the structure-aware features $f_i$, $i \in [1, N]$, the next step is to get the $w$ space of the StyleGAN backbone. Following previous StyleGAN models with the encoder for the degraded data [59], $f_N$ is first mapped into the $z$ space and then mapped to the $w$ space, as

$$
    w = M_w(z) = M_w(M_z(\mathcal{P}(f_N))),
$$

(5)

where $M_w$ and $M_z$ are the mapping functions and $\mathcal{P}$ is the pooling operation. The extracted $f_i$, $i \in [1, N]$ serve as the noise map to feed the structural information into the generator $G$, as shown in Fig. 3 (b).

3.3. Structure-Guided Enhancement Module

The structure modeling of the input low-light image $I$, as $I_s$, can be utilized to enhance the appearance predictions $I_a$. The enhancement can be interpreted from two aspects. First, using the structural information can enhance the image details, especially for the dark areas with low SNR, helping the generation of sharp and realistic results. The second effect is that the edge information help to distinguish different dark areas and build better relations among them.

Suppose the Structure-Guided Enhancement Module (SGEM) is denoted as $\mathcal{E}$, which can also be implemented as a simple U-Net. As shown in Fig. 3 (c), its input is the concatenation of $I_a$ and $I$ whose difference can provide the coarse direction for the enhancement on $I_a$. In this way, the enhancement can be denoted as

$$
    \hat{I} = \mathcal{E}(I_a \oplus I|I_s) + I_a,
$$

(6)

where $\oplus$ is the concatenation operation, and $I_s$ is the condition for the enhancement that is achieved by “Structure-Guided Feature Synthesis” in $\mathcal{E}$.

3.3.1 Structure-Guided Feature Synthesis

Suppose there are $K$ layers in the SGEM’s decoder, and the input of each layer is $d_j \in \mathbb{R}^{b \times p \times q}, j \in [1, K]$, where $b$, $p$, and $q$ are channel number, feature height, and width, respectively. In the $j$-th layer, we first resize the structure map $I_s$ into the same size of $d_j$, as $I_s^j$. To utilize the guidance of the structure map, we propose to generate spatially-varying kernels and normalization maps from $I_s^j$. The generation can be completed with Structure Guided Convolutions (SGC) and Structure Guided Normalizations (SGN), respectively. As shown in Fig. 3 (c), SGC produces different kernels for different locations. The feature map can be processed as

$$
    K_j = \mathcal{G}_{c,j}(I_s^j), \quad \hat{d}_j = d_j \ast K_j,
$$

(7)

where $\hat{d}_j \in \mathbb{R}^{b \times p \times q}$, “$\ast$” is the convolution operation, $K_j \in \mathbb{R}^{(b \times (k_h \times k_w)) \times p \times q}$ and $\mathcal{G}_{c,j}$ are the synthesized kernels ($k_h$ and $k_w$ are the kernel height and width) and the SGC in the $j$-th layer, respectively. SGN can predict different normalization maps for different locations to process the feature map, as

$$
    \alpha_j, \gamma_j = \mathcal{G}_{n,j}(I_s^j), \quad \hat{d}_j = IN(\hat{d}_j) \circ \alpha_j + \gamma_j,
$$

(8)

where $\alpha_j \in \mathbb{R}^{b \times p \times q}$ and $\gamma_j \in \mathbb{R}^{b \times p \times q}$ are the normalization parameters in the $j$-th layer, $\circ$ denotes the Hadamard product, $\hat{d}_j \in \mathbb{R}^{b \times p \times q}$, $IN$ is the instance normalization operation, and $\mathcal{G}_{n,j}$ is the SGN in the $j$-th layer. In the $j$-th layer, the feature can be enhanced as

$$
    d_{j+1} = C(d_j) + \hat{d}_j,
$$

(9)

where $C$ is the original convolutions in U-Net’s $j$-th layer.
where $L$ is the GAN loss function. The GAN loss is conducted by setting a discriminator $D$ effective in dealing with the drawback of the regression loss. One approach for supervising $S$ is employing the structure of the normal-light data as a ground truth and utilizing the regression loss. This strategy is adopted by existing edge prediction methods [33, 40].

3.4. Loss Function

Our framework is trained end-to-end, and the loss function can be divided into three parts.

**Loss for appearance modeling.** The loss function for the appearance modeling part $A$ is the reconstruction error between $I_a$ and the ground truth $I$. The loss is computed at both the pixel level and perceptual level, as

$$ L_a = ||I_a - I|| + ||\Phi(I_a) - \Phi(I)||, $$  \hspace{1cm} (10)  

where $\Phi()$ extracts features from the VGG network [39].

**Loss for structure modeling.** One approach for supervising $S$ is employing the structure of the normal-light data as the ground truth and utilizing the regression loss. This strategy is adopted by existing edge prediction methods [33, 40]. For the input of $I$, we assume the ground truth is the edge map extracted from $I$ (using the Canny edge detector [2]), as $I_s$. The regression loss is the binary cross-entropy, as

$$ L_s = -[I_s \log I_a + (1 - I_s) \log(1 - I_a)], \quad I_s = C(I), $$  \hspace{1cm} (11)  

where $C$ is the Canny detector. Moreover, the invisibility of content and the influence of noise in dark images would aggravate the ill-posed degree of estimating structures. Thus, it is difficult to detect structural details from the dark input picture with only the regression loss. We find the GAN loss is effective in dealing with the drawback of the regression loss. The GAN loss is conducted by setting a discriminator $D$, as

$$ L_g = \mathbb{E}_I \log(1 + \exp(-D(I_s))), \\ L_d = \mathbb{E}_I \log(1 + \exp(-D(I_s))) + \mathbb{E}_I \log(1 + \exp(+D(I_s)))). $$  \hspace{1cm} (12)  

where $L_g$ and $L_d$ are the loss functions for $S$ and $D$, respectively, and $\mathbb{E}$ is the mean operation.

**Loss for SGEM.** SGEM $E$ is to refine the appearance modeling, and the loss is also the reconstruction error, as

$$ L_m = ||\tilde{I} - I|| + ||\Phi(\tilde{I}) - \Phi(I)||. $$  \hspace{1cm} (13)  

**Overall loss function.** The overall framework is trained in end-to-end manner and the overall loss function is

$$ L = \lambda_1 L_a + \lambda_2 L_s + \lambda_3 L_g + \lambda_4 L_m, $$  \hspace{1cm} (14)  

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the loss weights.

4. Experiments

4.1. Implementation Details

We implement our framework in PyTorch [32]. Due to the limited computation resource, our framework is trained on 4 GPUs with RTX3090. For the loss minimization, we adopt the Adam [18] optimizer with momentum set to 0.9.

4.2. Datasets

Existing low-light enhancement datasets have different properties. SID [3] is a challenging dataset in both sRGB and RAW domains, and many dark areas need to be enhanced with structural guidance. For SID, each input sample is a pair of short- and long-exposure images, and low-light images have heavy noise since they were captured in extremely dark environments. We use the subset captured by the Sony camera of SID for experiments. Moreover, we also adopt LOL [62] dataset, which is divided into LOL-real and LOL-synthetic. Low-light images in LOL-real were collected from the real world; LOL-synthetic was created by synthesis.

4.3. Comparison on sRGB Domain

We compare our method with a rich collection of SOTA methods for low-light enhancement, including SID [3],

<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>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>16.71</td>
<td>17.01</td>
<td>18.02</td>
<td>18.34</td>
<td>20.08</td>
<td>20.84</td>
<td>16.44</td>
<td>20.11</td>
<td>20.67</td>
<td>18.55</td>
<td>20.33</td>
<td>18.91</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.591</td>
<td>0.578</td>
<td>0.600</td>
<td>0.578</td>
<td>0.598</td>
<td>0.605</td>
<td>0.596</td>
<td>0.602</td>
<td>0.587</td>
<td>0.617</td>
<td></td>
<td>0.617</td>
</tr>
<tr>
<td>PSNR</td>
<td>16.71</td>
<td>17.01</td>
<td>18.02</td>
<td>18.34</td>
<td>20.08</td>
<td>20.84</td>
<td>16.44</td>
<td>20.11</td>
<td>20.67</td>
<td>18.55</td>
<td>20.33</td>
<td>24.62</td>
</tr>
</tbody>
</table>
DeepUPE [44], KIND [69], DeepLF [30], FIDE [55], LPNet [22], MIR-Net [67], RF [19], 3DLUT [68], A3DLUT [46], Band [61], EG [13], Retinex [24], Sparse [62], UNIE [14], LCDR [43], LLFlow [49], DSN [70], RCTNet [17], UTNNet [71], SCI [28], URetinex [54], and SNR [56].

Quantitative analysis. We adopt PSNR and SSIM [51] for evaluation. A higher SSIM means more high-frequency details and structures in results. Tables 1 and 2 show the comparisons on LOL-real and LOL-synthetic. Our method surpasses all the baselines, and the improvement of SSIM is very evident thanks to the successful structure modeling and guidance. These numbers are obtained either from the respective papers or by running the respective public code. Table 3 shows the comparison on SID. It can be seen that although previous SOTA approaches, e.g., SNR [56], can achieve high PSNR, their corresponding SSIM values are not satisfying. Our framework again has the best performance over others and has obvious superiority to others by large margins on SSIM.

Qualitative analysis. We present visual samples in Fig. 4 to compare our method with the baseline that achieves the best performance on LOL-real and LOL-synthetic. Our result shows better visual quality with higher contrast, more accurate and clear details, more natural color consistency and brightness. Fig. 5 also shows visual comparisons on SID. While the input images in SID have evident noise and weak illumination, our method can produce more realistic results with fewer artifacts than others.

4.4. Comparison on RAW Domain

We compare our method with existing low-light image enhancement methods that are designed for RAW inputs. The baselines include CAN [5], DeepUPE [44], SID [3], EEMEFN [74], LLPackNet [20], FIDE [55], DID [29], SGN [8], DCE [9], RED [21], and ABF [6]. Compared with the enhancement on the sRGB domain, the RAW inputs have more information, leading to the enhancement with higher PSNR and SSIM.

The quantitative results are shown in Table 4. Our method still obtains the SOTA performance on the RAW domain in terms of both PSNR and SSIM. Our method results in the best SSIM that characterize the richness and sharpness of the images, since our approach formulates the effective structure modeling explicitly. The qualitative results in Fig. 5 also support the superiority of our framework in the RAW domain. As shown in the figure, these results also manifest that our method effectively enhances the image lightness and reveals details, while suppressing artifacts.

4.5. Ablation Study

We consider the following ablation settings by removing different components from our framework individually.

1. “Ours w/o \(A\)”: remove the module of \(A\), only the input image and the structure map are set as the input of \(E\).
2. “Ours w/o \(S\)”: delete the structural modeling module \(S\), and the framework is changed to the structure of two concatenated networks for appearance modeling.
3. “Ours w/o \(F\)”: replace the SAFE with the traditional encoder for the StyleGAN [59].
4. “Ours w/o \(G\)”: remove structure-guided feature synthesis in \(E\), and set the output of \(S\) as the input of \(E\).
5. “Ours w/o S.G.”: use the SOTA encoder-decoder-based edge prediction network [33] to implement $S$.

6. “Ours w/o GAN”: train $S$ without GAN loss.

We performed ablation studies on all three datasets on the sRGB domain. Table 5 summarizes the results. Compared with all ablation settings, our full setting yields higher PSNR and SSIM. Comparing “Ours w/o $A$” with “Ours” shows the necessity of conducting both the appearance and structure modeling in our framework. And the comparison between “Ours w/o $S$” and “Ours” demonstrates the effectiveness of the structure guidance to enhance appearance modeling. The results also show effects of “SAFE” (“Ours w/o $F$” vs. “Ours”), “structure-guided feature synthesis” (“Ours w/o $G$” vs. “Ours”), and the GAN loss in structure modeling (“Ours w/o GAN” vs. “Ours”). Moreover, comparing “Ours w/o S.G.” with “Ours” proves the superiority of our $S$ over existing encoder-decoder-based edge detection networks on low-light images.

Furthermore, to demonstrate the robustness of our framework, we conduct the evaluation setting where extra noise (the Gaussian distribution with mean as 0, variance ranges from 30 to 50) is added in the input image. The results are displayed in Table 5 as “Ours with noise”, which is close to the situation without noise, showing the robustness of our framework towards perturbations.

### 4.6. Evaluation for Structural Modeling

In this section, we conduct ablation studies for the structure modeling. Since we set the edge map of the normal-light image as the ground truth, we use the cross-entropy (CE) and the $L_2$ distance between the prediction and the ground truth as the metric. We compare the ablation settings that also employ the structure modeling, i.e., “Ours w/o GAN”, “Ours w/o S.G.”, and “Ours w/o $F$”. The results are reported in Table 6. Compared with all ablation settings, our full setting achieves the most accurate structure modeling from the low-light images. By comparing “Ours w/o $F$” and “Ours w/o S.G.” with “Ours”, we can demonstrate the superiority of our $S$ in the structure modeling to other alternatives. By comparing “Ours” with “Ours w/o
### 4.7. User Study

We conduct a large-scale user study with 100 participants for human subjective evaluation, and the five strongest baselines are chosen by mean PSNR on SID and LOL. All methods are trained on SID, and we evaluate their performances on the real-captured low-light photos (a total of 50 images) from an iPhone 8. The capturing environments are various, including indoor and outdoor scenes. Following the settings in [56], the evaluation is completed via user ratings on the six questions shown in Fig. 7, where scores range from 1 (worst) to 5 (best). Fig. 7 exhibits the rating distributions of different methods. We can see that our framework receives more high and fewer low ratings, showing the superior perception of our results.

### 4.8. Structure Modeling with More Data

All the above results (ours and baselines) are obtained from the networks trained without extra data. The collection of paired low-light and normal-light data, i.e., \( \hat{I} \) and \( I \), is difficult and expensive, while the paired data to supervise the structure modeling \( S \) is trivial since the ground truth \( \hat{I} \) can be obtained with the edge detection algorithm from \( \hat{I} \). Therefore, we can easily obtain extra data pair for \( S \) from another dataset containing normal-light data. Following [60], we adopt AVA dataset [31] to acquire the data pair to fine-tune \( S \). The result can be improved again by synthesizing better structure for very dark areas, as shown in Table 7 (Ours with E.D.).

### 4.9. Ours with Other Structural Representations

In this work, we adopt the edge for the structure modeling since the edge is a general structural representation for various scenes. Our framework is also applicable to other structural representations, e.g., the segmentation and depth maps. We adopt the DPT model [35], which is trained on ADE20K [72] and MIX 6 [35], to extract the normal-light data’s segmentation maps and depth maps. And they are employed as the ground truth to train \( S \). As shown in Table 7, all these results (“Ours with Seg.” and “Ours with Dep.”) are better than the baselines in Tables 1, 2, 3, demonstrating the effectiveness of our structure modeling and guidance strategy.

### 5. Conclusion

In this paper, we propose a new framework to conduct structure modeling and employ the restored structure maps to enhance the appearance modeling results. Different from existing methods, the structure modeling in our framework is undertaken via a modified generative model with structural features and is trained with the GAN loss. We further design a novel structure-guided enhancement module for appearance enhancement with structure-guided feature synthesis layers. Extensive experiments on sRGB and RAW domains demonstrate the effectiveness of our framework.

**Acknowledgments.** This work is partially supported by Shenzhen Science and Technology Program KQTD20210811090149095.
References


[29] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019. 6


[31] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 8


[33] Mengyang Pu, Yaping Huang, Yuming Liu, Qingji Guan, and Haibin Ling. EDTER: Edge detection with transformer. In *CVPR*, 2022. 5, 7


