# Low-Light Image Enhancement via Structure Modeling and Guidance

Xiaogang Xu[1]    Ruixing Wang[2]    Jiangbo Lu[3*]

[1] Zhejiang Lab    [2] Honor Device Co., Ltd.    [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 [11, 27, 50], and also benefits many downstream vision tasks, e.g., nighttime detection [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 [7, 10, 45, 63], and some recent works also account for noise in images [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
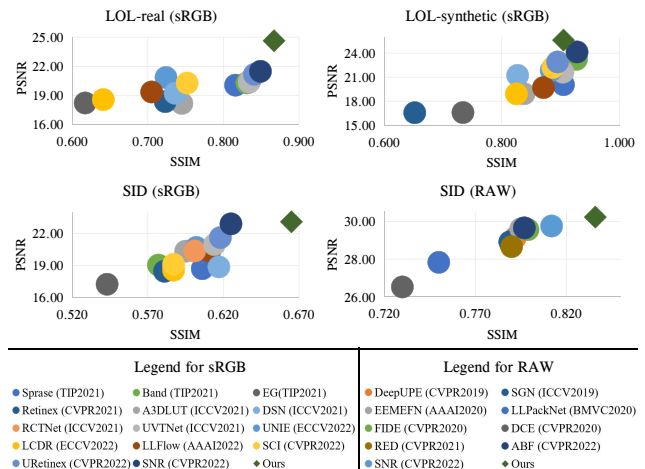
---

*Corresponding author.



Figure 1. Our method *consistently* achieves SOTA performance on *different sRGB/RAW datasets* with the same network architecture.

resulting in blurry outcomes and low SSIM [51] values, as shown in Fig. 2. Some works [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 [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 [34, 74].

In this paper, we propose to utilize a generative model $\mathcal{S}$ 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 $\mathcal{A}$) 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,

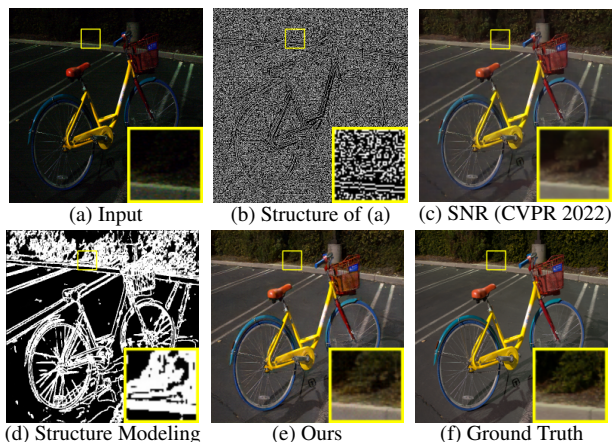| (a) Input | (b) Structure of (a) | (c) SNR (CVPR 2022) |
| (d) Structure Modeling | (e) Ours | (f) Ground Truth |

Figure 2. A challenging low-light frame (a), from SID-sRGB [3], enhanced by a SOTA method (c) and our method (e). Our method can synthesize the structure map (d) from the input image, leading to clearer details, more distinct contrast, and more vivid color. Although (c) has high PSNR as 28.17, its SSIM is low as 0.75. Ours achieves high scores for both dB and SSIM, as 28.60dB and 0.80.

the proposed generative model $\mathcal{S}$ has two significant modifications. First, we notice the impact of providing structure-aware descriptors into both the encoder and decoder of $\mathcal{S}$, 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 $\mathcal{S}$ 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.

$\mathcal{S}$, $\mathcal{A}$, 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/gradients [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, can not 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.
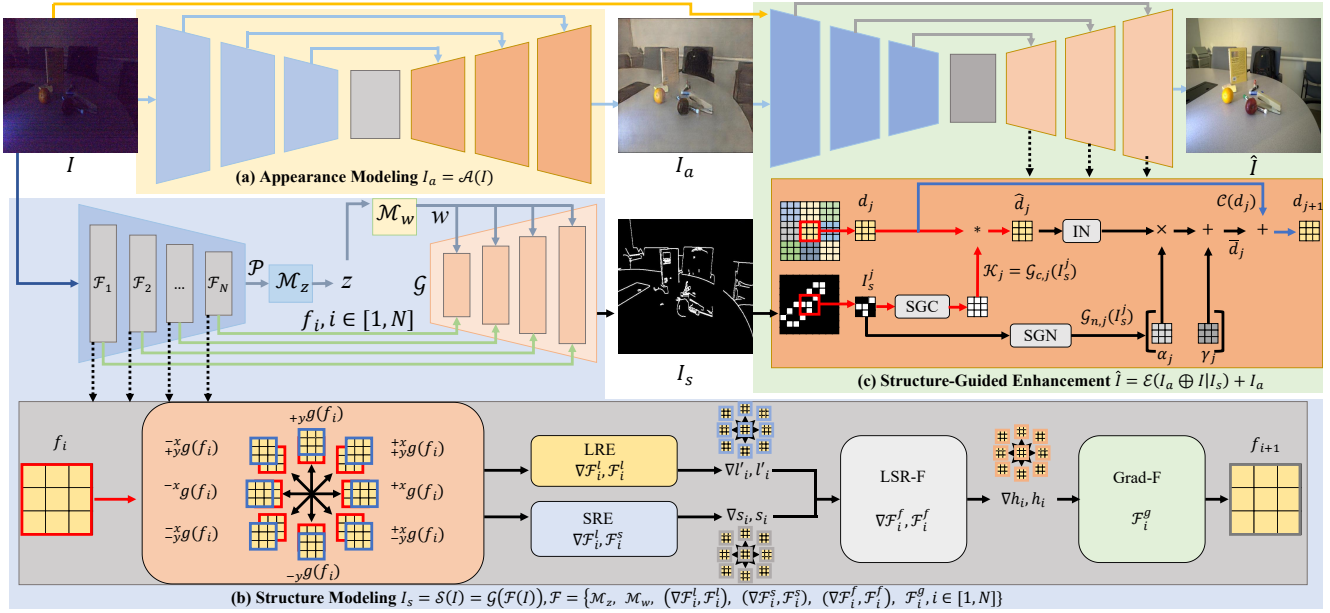
Figure 3. The overview of our framework with explicit appearance modeling $\mathcal{A}$, structure modeling $\mathcal{S}$, and SGEM $\mathcal{E}$. The supervision for $I_a$ and $\widehat{I}$ is the normal-light image $\bar{I}$, and for $I_s$ is the edge $\bar{I}_s$ extracted from $\bar{I}$. The overall framework can be trained end-to-end.

## 3. Method

The overview of our framework can be viewed in Fig. 3. The proposed framework conducts explicit modeling for both appearance $\mathcal{A}$ as well as structure $\mathcal{S}$, 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 $\mathcal{E}$, 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 [3] as the network for appearance modeling, as shown in Fig. 3 (a), where $I_a = \mathcal{A}(I)$. 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 [15], can be modified to be a strong structure estimator $\mathcal{S}$ for low-light images as shown in Fig. 3 (b). Equipped with both the structure regression loss and the GAN loss, $\mathcal{S}$ can accurately predict the effective structure maps that are beneficial for appearance enhancement. In this paper, we use the backbone of StyleGAN to formulate $\mathcal{S}$. Given the input image $I$, we can obtain its structure map $\mathcal{S}$ as

$$I_s = \mathcal{S}(I) = \mathcal{G}(\mathcal{F}(I)), \tag{1}$$

where $\mathcal{F}$ is the encoder part, as "Structure-Aware Feature Extractor (SAFE)" that contains $N$ layers and two mapping functions, and $\mathcal{G}$ is decoder part, as "Structure-Aware Generator (SAG)". $\mathcal{F}$ and $\mathcal{G}$ 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 [37, 42, 47], e.g., CNN. However, as indicated in [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, $\mathcal{F}_1, ..., \mathcal{F}_N$, 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 $\mathcal{F}_l$ and $\mathcal{F}_s$, the feature extraction process can be written as

$$
\begin{aligned}
l_i &= \mathcal{F}_i^l(f_i), & s_i &= \mathcal{F}_i^s(f_i), \\
\triangledown l_i &= \triangledown \mathcal{F}_i^l(\triangledown g(f_i)), & \triangledown s_i &= \triangledown \mathcal{F}_i^s(\triangledown g(f_i)),
\end{aligned} \tag{2}
$$

where $\triangledown \in \{{}^{+x}, {}^{-x}, {}_{+y}, {}_{-y}, {}^{+x}_{+y}, {}^{-x}_{+y}, {}^{+x}_{-y}, {}^{-x}_{-y}\}$, $l_i$ 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 $\mathcal{F}_i^f$, then the spatially-varing extracted features for $f$ and $g$ can be formulated as

$$
h_i = \mathcal{F}_i^f(l_i, s_i), \quad \triangledown h_i = \triangledown \mathcal{F}_i^f(\triangledown l_i, \triangledown s_i). \tag{3}
$$

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

$$
\begin{aligned}
f_{i+1} = \mathcal{F}_i^g(h_i, {}^{+x}h_i, {}^{-x}h_i, {}_{-y}h_i, {}_{+y}h_i, \\
{}^{+x}_{+y}h_i, {}^{-x}_{+y}h_i, {}^{+x}_{-y}h_i, {}^{-x}_{-y}h_i),
\end{aligned} \tag{4}
$$

where $\mathcal{F}_i^g$ 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

$$
w = \mathcal{M}_w(z) = \mathcal{M}_w(\mathcal{M}_z(\mathcal{P}(f_N))), \tag{5}
$$

where $\mathcal{M}_w$ and $\mathcal{M}_z$ 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 $\mathcal{G}$, 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

$$
\widehat{I} = \mathcal{E}(I_a \oplus I | I_s) + I_a, \tag{6}
$$

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

$$
\mathcal{K}_j = \mathcal{G}_{c,j}(I_s^j), \quad \widehat{d}_j = d_j * \mathcal{K}_j, \tag{7}
$$

where $\widehat{d}_j \in \mathbb{R}^{b \times p \times q}$, "$*$" is the convolution operation, $\mathcal{K}_j \in \mathbb{R}^{(b \times (k_h \times k_w)) \times p \times q}$ 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

$$
\alpha_j, \gamma_j = \mathcal{G}_{n,j}(I_s^j), \quad \bar{d}_j = IN(\widehat{d}_j) \circ \alpha_j + \gamma_j, \tag{8}
$$

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, $\bar{d}_j \in \mathbb{R}^{b \times p \times q}$, $IN$ 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

$$
d_{j+1} = \mathcal{C}(d_j) + \bar{d}_j, \tag{9}
$$

where $\mathcal{C}$ is the original convolutions in U-Net's $j$-th layer.

| Methods | SID [3] | DeepUPE [44] | KIND [69] | DeepLPF [30] | FIDE [55] | LPNet [22] | MIR-Net [67] | RF [19] | 3DLUT [68] | UNIE [14] | LCDR [43] | LLFlow [49] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PSNR | 13.24 | 13.27 | 14.74 | 14.10 | 16.85 | 17.80 | 20.02 | 14.05 | 17.59 | 20.85 | 18.57 | 19.36 |
| SSIM | 0.442 | 0.452 | 0.641 | 0.480 | 0.678 | 0.792 | 0.820 | 0.458 | 0.721 | 0.724 | 0.641 | 0.705 |
| Methods | A3DLUT [46] | Band [61] | EG [13] | Retinex [24] | Sparse [62] | DSN [70] | RCTNet [17] | UTVNet [71] | SCI [28] | URetinex [54] | SNR [56] | Ours |
| PSNR | 18.19 | 20.29 | 18.23 | 18.37 | 20.06 | 19.23 | 20.51 | 20.37 | 20.28 | 21.16 | 21.48 | **24.62** |
| SSIM | 0.745 | 0.831 | 0.617 | 0.723 | 0.815 | 0.736 | 0.831 | 0.834 | 0.752 | 0.840 | 0.849 | **0.867** |

Table 1. Quantitative comparison on the LOL-real dataset.

| Methods | SID [3] | DeepUPE [44] | KIND [69] | DeepLPF [30] | FIDE [55] | LPNet [22] | MIR-Net [67] | RF [19] | 3DLUT [68] | UNIE [14] | LCDR [43] | LLFlow [49] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PSNR | 15.04 | 15.08 | 13.29 | 16.02 | 15.20 | 19.51 | 21.94 | 15.97 | 18.04 | 21.84 | 18.91 | 19.69 |
| SSIM | 0.610 | 0.623 | 0.578 | 0.587 | 0.612 | 0.846 | 0.876 | 0.632 | 0.800 | 0.884 | 0.825 | 0.871 |
| Methods | A3DLUT [46] | Band [61] | EG [13] | Retinex [24] | Sparse [62] | DSN [70] | RCTNet [17] | UTVNet [71] | SCI [28] | URetinex [54] | SNR [56] | Ours |
| PSNR | 18.92 | 23.22 | 16.57 | 16.55 | 22.05 | 21.22 | 22.44 | 21.62 | 22.20 | 22.89 | 24.14 | **25.62** |
| SSIM | 0.838 | 0.927 | 0.734 | 0.652 | 0.905 | 0.827 | 0.891 | 0.904 | 0.887 | 0.895 | **0.928** | 0.905 |

Table 2. Quantitative comparison on the LOL-synthetic dataset.

| Methods | SID [3] | DeepUPE [44] | KIND [69] | DeepLPF [30] | FIDE [55] | LPNet [22] | MIR-Net [67] | RF [19] | 3DLUT [68] | UNIE [14] | LCDR [43] | LLFlow [49] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PSNR | 16.97 | 17.01 | 18.02 | 18.07 | 18.34 | 20.08 | 20.84 | 16.44 | 20.11 | 20.67 | 18.55 | 20.33 |
| SSIM | 0.591 | 0.604 | 0.583 | 0.600 | 0.578 | 0.598 | 0.605 | 0.596 | 0.592 | 0.602 | 0.587 | 0.611 |
| Methods | A3DLUT [46] | Band [61] | EG [13] | Retinex [24] | Sparse [62] | DSN [70] | RCTNet [17] | UTVNet [71] | SCI [28] | URetinex [54] | SNR [56] | Ours |
| PSNR | 20.32 | 19.02 | 17.23 | 18.44 | 18.68 | 18.85 | 20.34 | 20.93 | 19.09 | 21.56 | 22.87 | **23.18** |
| SSIM | 0.595 | 0.577 | 0.543 | 0.581 | 0.606 | 0.617 | 0.601 | 0.614 | 0.585 | 0.619 | 0.625 | **0.664** |

Table 3. Quantitative comparison on the SID dataset (sRGB domain).

## 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 $\mathcal{A}$ is the reconstruction error between $I_a$ and the ground truth $\bar{I}$. The loss is computed at both the pixel level and perceptual level, as

$$\mathcal{L}_a = \|I_a - \bar{I}\| + \|\Phi(I_a) - \Phi(\bar{I})\|, \qquad (10)$$

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

**Loss for structure modeling.** One approach for supervising $\mathcal{S}$ 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 $\bar{I}$ (using the Canny edge detector [2]), as $\bar{I}_s$. The regression loss is the binary cross-entropy, as

$$\mathcal{L}_s = -[\bar{I}_s \log I_s + (1 - \bar{I}_s) \log(1 - I_s)], \ \bar{I}_s = C(\bar{I}), \quad (11)$$

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 $\mathcal{D}$, as

$$\mathcal{L}_g = \mathbb{E}_I(\log(1 + \exp(-\mathcal{D}(I_s)))),$$
$$\mathbb{E}_I \log(1 + \exp(-\mathcal{D}(\mathcal{S}(I)))), \mathcal{L}_d = \mathbb{E}_I(\log(1 + \exp(-\mathcal{D}(\bar{I}_s)))) +$$
$$\mathbb{E}_I(\log(1 + \exp(+\mathcal{D}(I_s)))), \qquad (12)$$

where $\mathcal{L}_g$ and $\mathcal{L}_d$ are the loss functions for $\mathcal{S}$ and $\mathcal{D}$, respectively, and $\mathbb{E}$ is the mean operation.

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

$$\mathcal{L}_m = \|\widehat{I} - \bar{I}\| + \|\Phi(\widehat{I}) - \Phi(\bar{I})\|. \qquad (13)$$

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

$$\mathcal{L} = \lambda_1 \mathcal{L}_a + \lambda_2 \mathcal{L}_s + \lambda_3 \mathcal{L}_g + \lambda_4 \mathcal{L}_m, \qquad (14)$$

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],
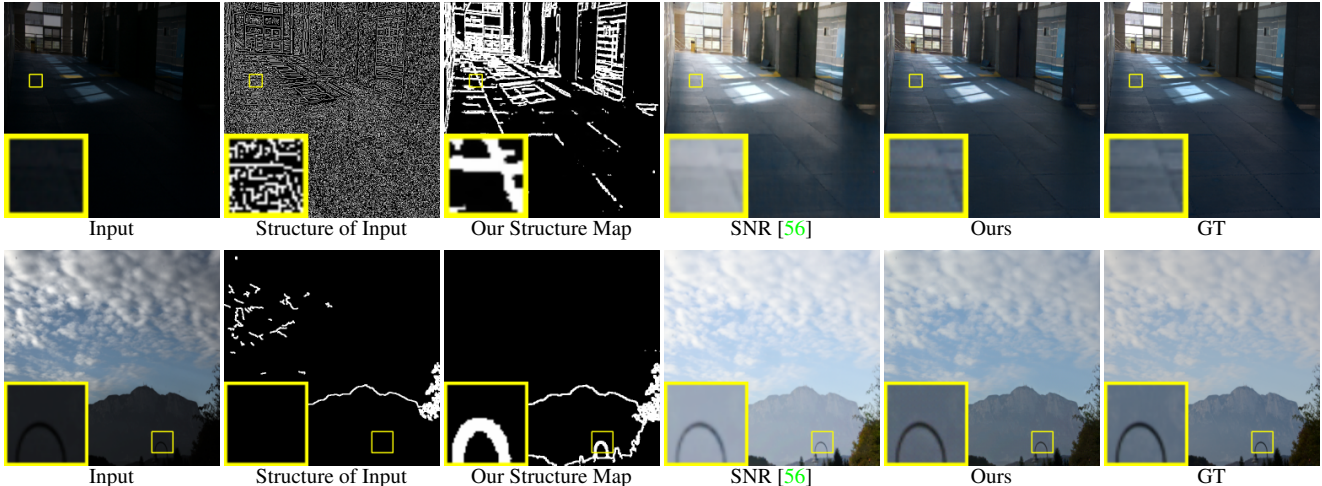
Figure 4. Visual comparisons on LOL-real (top) and LOL-synthetic (bottom). "Ours" have less noise and clearer visibility.

DeepUPE [44], KIND [69], DeepLPF [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], UTVNet [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

| Methods | DeepUPE [44] | SID [3] | EEMEFN [74] | DCE [9] |
|---------|--------------|---------|-------------|---------|
| PSNR | 29.13 | 28.88 | 29.60 | 26.53 |
| SSIM | 0.792 | 0.787 | 0.795 | 0.730 |
| Methods | LLPackNet [20] | FIDE [55] | DID [29] | SGN [8] |
| PSNR | 27.83 | 29.56 | 28.41 | 28.91 |
| SSIM | 0.750 | 0.799 | 0.780 | 0.789 |
| Methods | RED [21] | ABF [6] | SNR [56] | Ours |
| PSNR | 28.66 | 29.65 | 29.75 | **30.17** |
| SSIM | 0.790 | 0.797 | 0.812 | **0.834** |

Table 4. Quantitative comparison on SID's RAW domain.

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 $\mathcal{A}$": remove the module of $\mathcal{A}$, only the input image and the structure map are set as the input of $\mathcal{E}$.

2. "Ours w/o $\mathcal{S}$": delete the structural modeling module $\mathcal{S}$, and the framework is changed to the structure of two concatenated networks for appearance modeling.

3. "Ours w/o $\mathcal{F}$": replace the SAFE with the traditional encoder for the StyleGAN [59].

4. "Ours w/o $\mathcal{G}$": remove structure-guided feature synthesis in $\mathcal{E}$, and set the output of $\mathcal{S}$ as the input of $\mathcal{E}$.
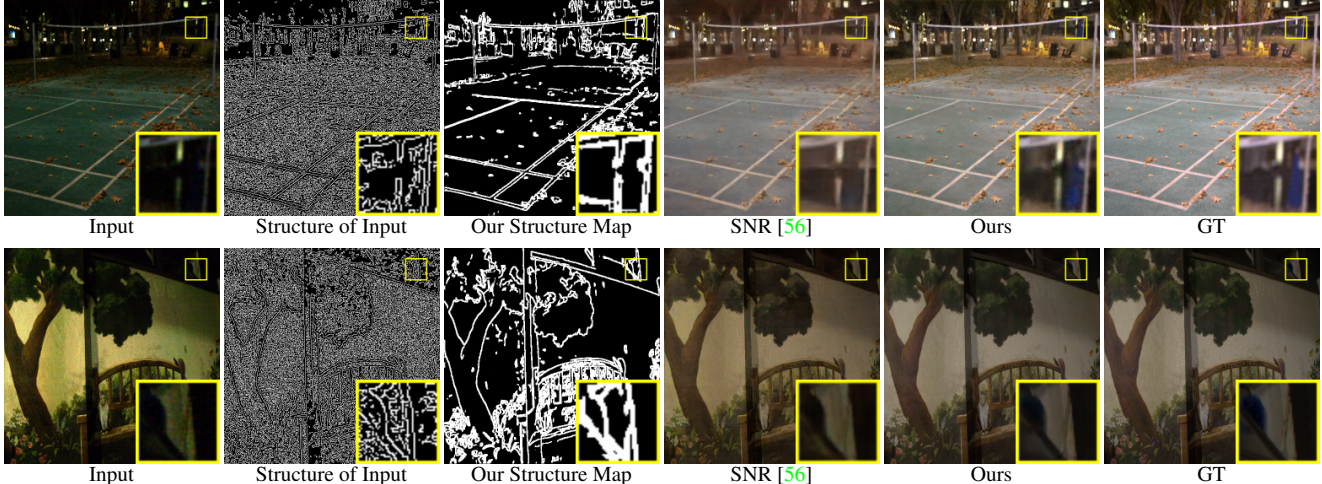
Figure 5. Visual comparisons on SID's sRGB (top) and RAW (bottom) domains. "Ours" have less noise and clearer visual details.

5. "Ours w/o S.G.": use the SOTA encoder-decoder-based edge prediction network [33] to implement $\mathcal{S}$.

6. "Ours w/o GAN": train $\mathcal{S}$ 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 $\mathcal{A}$" with "Ours" shows the necessity of conducting both the appearance and structure modeling in our framework. And the comparison between "Ours w/o $\mathcal{S}$" and "Ours" demonstrates the effectiveness of the structure guidance to enhance appearance modeling. The results also show effects of "SAFE" ("Ours w/o $\mathcal{F}$" 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 $\mathcal{S}$ 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 $\mathcal{F}$". The results are reported in Table 6. Compared with all ablation

|  | LOL-real | | LOL-synthetic | | SID | |
|---|---|---|---|---|---|---|
| Methods | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Ours w/o $\mathcal{A}$ | 20.17 | 0.801 | 23.59 | 0.879 | 22.59 | 0.639 |
| Ours w/o $\mathcal{S}$ | 18.14 | 0.773 | 21.20 | 0.881 | 20.47 | 0.623 |
| Ours w/o $\mathcal{F}$ | 20.21 | 0.812 | 23.05 | 0.888 | 22.35 | 0.635 |
| Ours w/o $\mathcal{G}$ | 19.39 | 0.784 | 21.71 | 0.868 | 21.15 | 0.629 |
| Ours w/o S.G. | 20.73 | 0.820 | 23.30 | 0.898 | 22.50 | 0.632 |
| Ours w/o GAN | 21.28 | 0.812 | 23.17 | 0.883 | 22.14 | 0.642 |
| Results with $\mathcal{A}$ | 18.99 | 0.715 | 21.76 | 0.863 | 19.34 | 0.556 |
| Ours with noise | 24.15 | 0.832 | 24.07 | 0.880 | 22.86 | 0.648 |
| Ours | **24.62** | **0.867** | **25.62** | **0.905** | **23.18** | **0.664** |

Table 5. Results of the ablation study in Sec. 4.5.

|  | LOL-real | | LOL-synthetic | | SID | |
|---|---|---|---|---|---|---|
| Methods | CE | $L_2$ | CE | $L_2$ | CE | $L_2$ |
| Ours w/o GAN | 0.2581 | 0.3650 | 0.2144 | 0.3936 | 0.5335 | 0.5034 |
| Ours w/o S.G. | 0.2923 | 0.3805 | 0.2133 | 0.3833 | 0.5035 | 0.5405 |
| Ours w/o $\mathcal{F}$ | 0.3070 | 0.3553 | 0.2795 | 0.3675 | 0.5351 | 0.4905 |
| Ours | **0.2130** | **0.3042** | **0.2072** | **0.3032** | **0.4352** | **0.4541** |

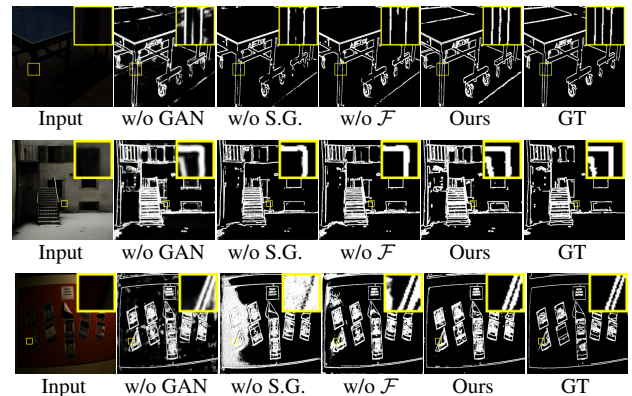Table 6. Results of the ablation study in Sec. 4.6.


Figure 6. Visual comparisons for structure modeling on LOL-real, LOL-synthetic, and SID datasets (top to bottom).

settings, our full setting achieves the most accurate structure modeling from the low-light images. By comparing "Ours w/o $\mathcal{F}$" 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
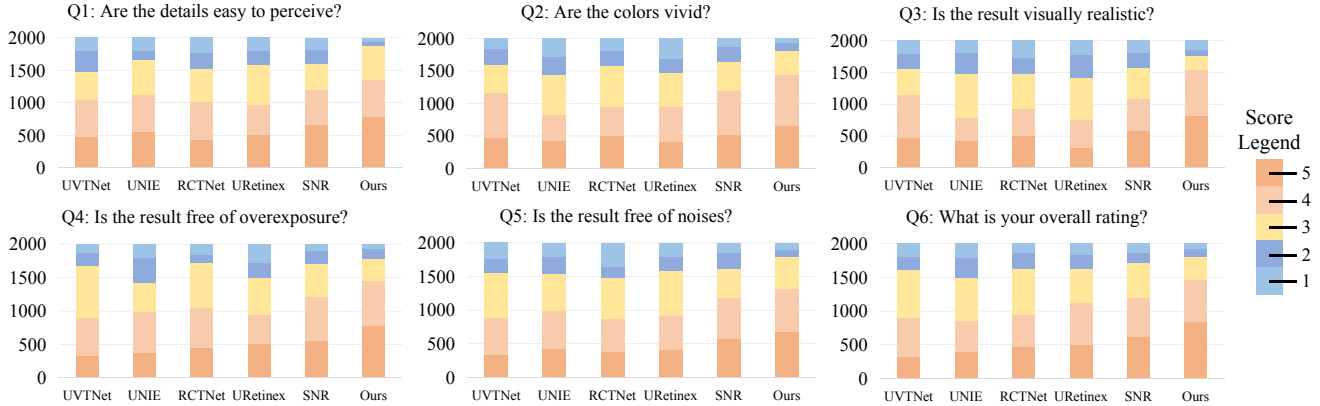
Figure 7. The results of the user study. The ordinate axis denotes the rating frequency received from all participants.

| Methods | LOL-real | | LOL-synthetic | | SID | |
|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Ours with E.D. | 25.57 | 0.885 | 26.48 | 0.951 | 24.43 | 0.707 |
| Ours with Seg. | 24.18 | 0.846 | 25.27 | 0.903 | 23.10 | 0.659 |
| Ours with Dep. | 23.89 | 0.830 | 24.91 | 0.902 | 22.95 | 0.640 |
| Ours | 24.62 | 0.867 | 25.62 | 0.905 | 23.18 | 0.664 |

Table 7. Results of experiments in Sec. 4.8 and 4.9.

GAN", we show the effects of the GAN loss in the structure modeling from low-light images. The visual comparisons are also provided in Fig. 6. We can see that the edge predictions' uncertainty will be increased without the GAN loss, causing the artifacts in the edge prediction. When compared with the results of "Ours w/o $\mathcal{F}$" and "Ours w/o S.G.", our results have more edge details and fewer artifacts.

### 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., $I$ and $\bar{I}$, is difficult and expensive, while the paired data to supervise the structure modeling $\mathcal{S}$ is trivial since the ground truth $\bar{I}_s$ can be obtained with the edge detection algorithm from $\bar{I}$. Therefore, we can easily obtain extra data pair for $\mathcal{S}$ from another dataset containing normal-light data. Following [60], we adopt AVA dataset [31] to acquire the data

pair to finetune $\mathcal{S}$. 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 $\mathcal{S}$. 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.

**Limitation.** The structure generation with our model in extremely dark areas (with almost no information) could have artifacts when only trained with existing low-light datasets. This issue can be alleviated by training with large-scale data as indicated in Sec. 4.8, requiring more computation resources and a long training time.

# References

[1] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE TIP*, 2018. 2

[2] John Canny. A computational approach to edge detection. *IEEE TPAMI*, 1986. 5

[3] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 2, 3, 5, 6

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 4

[5] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. In *ICCV*, 2017. 6

[6] Xingbo Dong, Wanyan Xu, Zhihui Miao, Lan Ma, Chao Zhang, Jiewen Yang, Zhe Jin, Andrew Beng Jin Teoh, and Jiajun Shen. Abandoning the Bayer-Filter to see in the dark. In *CVPR*, 2022. 6

[7] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, 2016. 1

[8] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *ICCV*, 2019. 6

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

[10] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 2016. 1

[11] Shijie Hao, Xu Han, Yanrong Guo, Xin Xu, and Meng Wang. Low-light image enhancement with semi-decoupled decomposition. *IEEE TMM*, 2020. 1, 2

[12] Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. GCFSR: a generative and controllable face super resolution method without facial and GAN priors. In *CVPR*, 2022. 2

[13] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE TIP*, 2021. 2, 5, 6

[14] Yeying Jin, Wenhan Yang, and Robby T Tan. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *ECCV*, 2022. 5, 6

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3

[16] Eungyeup Kim, Sanghyeon Lee, Jeonghoon Park, Somi Choi, Choonghyun Seo, and Jaegul Choo. Deep edge-aware interactive colorization against color-bleeding effects. In *ICCV*, 2021. 2

[17] Hanul Kim, Su-Min Choi, Chang-Su Kim, and Yeong Jun Koh. Representative color transform for image enhancement. In *ICCV*, 2021. 2, 5, 6

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[19] Satoshi Kosugi and Toshihiko Yamasaki. Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. In *AAAI*, 2020. 5, 6

[20] Mohit Lamba, Atul Balaji, and Kaushik Mitra. Towards fast and light-weight restoration of dark images. In *BMVC*, 2020. 6

[21] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *CVPR*, 2021. 6

[22] Jiaqian Li, Juncheng Li, Faming Fang, Fang Li, and Guixu Zhang. Luminance-aware pyramid network for low-light image enhancement. *IEEE TMM*, 2020. 5, 6

[23] Dong Liang, Ling Li, Mingqiang Wei, Shuo Yang, Liyan Zhang, Wenhan Yang, Yun Du, and Huiyu Zhou. Semantically contrastive learning for low-light image enhancement. In *AAAI*, 2022. 2

[24] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, 2021. 1, 2, 5, 6

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4

[26] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. LL-Net: A deep autoencoder approach to natural low-light image enhancement. *PR*, 2017. 2

[27] Yucheng Lu and Seung-Won Jung. Progressive joint low-light enhancement and noise removal for raw images. *IEEE TIP*, 2022. 1

[28] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. 1, 5, 6

[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

[30] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. DeepLPF: Deep local parametric filters for image enhancement. In *CVPR*, 2020. 5, 6

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

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

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

[34] Deepanshu Rana, Kanishk Jayant Lal, and Anil Singh Parihar. Edge guided low-light image enhancement. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021. 1, 2

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 8

[36] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE TIP*, 2019. 2

[37] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *CVPR*, 2021. 3

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014. 5

[40] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *ICCV*, 2021. 5

[41] Masayuki Tanaka, Takashi Shibata, and Masatoshi Okutomi. Gradient-based low-light image enhancement. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*, 2019. 2

[42] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM TOG*, 2021. 3

[43] Haoyuan Wang, Ke Xu, and Rynson WH Lau. Local color distributions prior for image enhancement. In *ECCV*, 2022. 5, 6

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

[45] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE TIP*, 2013. 1

[46] Tao Wang, Yong Li, Jingyang Peng, Yipeng Ma, Xian Wang, Fenglong Song, and Youliang Yan. Real-time image enhancer via learnable spatial-aware 3D lookup tables. In *ICCV*, 2021. 2, 5, 6

[47] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity GAN inversion for image attribute editing. In *CVPR*, 2022. 3

[48] Wenjing Wang, Wenhan Yang, and Jiaying Liu. Hla-face: Joint high-low adaptation for low light face detection. In *CVPR*, 2021. 1

[49] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *AAAI*, 2022. 5, 6

[50] Yun-Fei Wang, He-Ming Liu, and Zhao-Wang Fu. Low-light image enhancement via the absorption light scattering model. *IEEE TIP*, 2019. 1

[51] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 1, 6

[52] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general U-Shaped transformer for image restoration. In *CVPR*, 2022. 4

[53] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. In *ICCV*, 2021. 4

[54] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. URetinex-Net: Retinex-Based deep unfolding network for low-light image enhancement. In *CVPR*, 2022. 5, 6

[55] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*, 2020. 1, 2, 5, 6

[56] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-Aware low-light image enhancement. In *CVPR*, 2022. 2, 3, 5, 6, 7, 8

[57] Jianzhou Yan, Stephen Lin, Bing Kang Sing, and Xiaoou Tang. A learning-to-rank approach for image color enhancement. In *CVPR*, 2014. 2

[58] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM TOG*, 2016. 2

[59] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. GAN prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 2, 4, 6

[60] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, 2020. 8

[61] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality. *IEEE TIP*, 2021. 2, 5, 6

[62] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep Retinex network for robust low-light image enhancement. *IEEE TIP*, 2021. 2, 5, 6

[63] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new low-light image enhancement algorithm using camera response model. In *ICCV*, 2017. 1

[64] Jun Yu, Xinlong Hao, and Peng He. Single-stage face detection under extremely low-light conditions. In *ICCV*, 2021. 1

[65] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *ICCV*, 2021. 4

[66] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet. In *ICCV*, 2021. 4

[67] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020. 2, 5, 6

[68] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. *IEEE TPAMI*, 2020. 2, 5, 6

[69] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, 2019. 5, 6

[70] Lin Zhao, Shao-Ping Lu, Tao Chen, Zhenglu Yang, and Ariel Shamir. Deep symmetric network for underexposed image enhancement with recurrent attentional learning. In *ICCV*, 2021. 2, 5, 6

[71] Chuanjun Zheng, Daming Shi, and Wentian Shi. Adaptive unfolding total variation network for low-light image enhancement. In *ICCV*, 2021. 2, 5, 6

[72] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 8

[73] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. Blind face restoration via integrating face shape and generative priors. In *CVPR*, 2022. 2

[74] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *AAAI*, 2020. 1, 2, 6