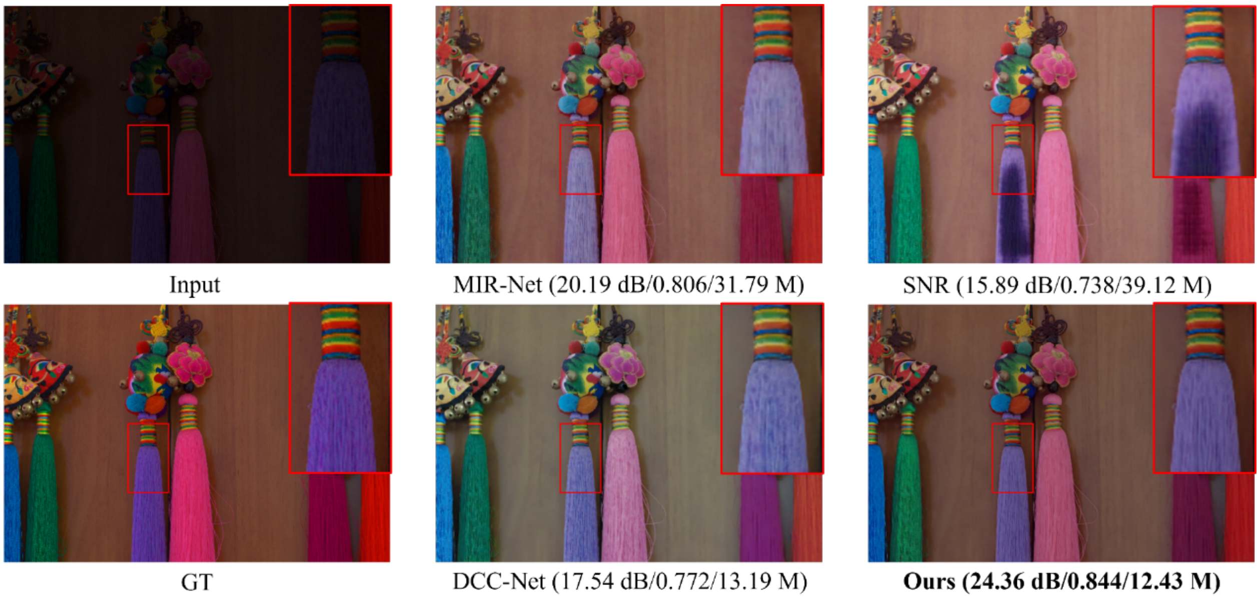
Low-Light Image Enhancement with Multi-stage Residue Quantization and Brightness-aware Attention

Yunlong Liu[1,*] Tao Huang[1,*] Weisheng Dong[1,†] Fangfang Wu[1] Xin Li[2] Guangming Shi[1]
[1] Xidian University [2] University at Albany
liuyunlong@stu.xidian.edu.cn thuang_666@stu.xidian.edu.cn wsdong@mail.xidian.edu.cn
wufangfang@xidian.edu.cn xli48@albany.edu gmshi_xidian@163.com



Input

MIR-Net (20.19 dB/0.806/31.79 M)

SNR (15.89 dB/0.738/39.12 M)

GT

DCC-Net (17.54 dB/0.772/13.19 M)

Ours (24.36 dB/0.844/12.43 M)

Figure 1: Qualitative and quantitative (i.e., PSNR / SSIM / parameters) comparisons of multiple state-of-the-art low-light image enhancement methods, including MIR-Net [45], DCC-Net [47], SNR [41], and our proposed method. Apparently, our proposed method has better brightness and fewer parameters than other competing methods.

Abstract

Low-light image enhancement (LLIE) aims to recover illumination and improve the visibility of low-light images. Conventional LLIE methods often produce poor results because they neglect the effect of noise interference. Deep learning-based LLIE methods focus on learning a mapping function between low-light images and normal-light images that outperforms conventional LLIE methods. However, most deep learning-based LLIE methods cannot yet fully exploit the guidance of auxiliary priors provided by normal-light images in the training dataset. In this paper, we propose a brightness-aware network with normal-light priors based on brightness-aware attention and residual-quantized codebook. To achieve a more natural and realistic enhancement, we design a query module to obtain more reliable normal-light features and fuse them with low-light features by a fusion branch. In addition, we propose a brightness-aware attention module to further improve the robustness of the network to the brightness. Extensive experimental results on both real-captured and synthetic data show that our method outperforms existing state-of-the-art methods.

1. Introduction

Low-light images suffer from extremely limited visibility and noise interference, which often have serious effects on the performance of many downstream tasks [6, 8, 29,

* Equal Contribution, † Corresponding Author

30]. Professional photographers can work with exposure time, aperture, and ISO settings to capture more information related to the image. However, in the meantime, motion blur degradation and noise amplification are often inevitable. Therefore, computational photography methods known as low-light image enhancement (LLIE) [9, 39, 43] have received increasing attention in recent years. Conventional LLIE methods, including histogram equalization methods [15, 25, 26], gamma correction methods [12, 27], and Retinex-based methods [14, 19, 39, 43], generally neglect noise degeneration, resulting in unsatisfactory performance on real low-light images. LLIE methods based on deep learning aim to learn a mapping function between low-light images and normal-light images and outperform conventional LLIE methods by a large margin [41, 45, 47]. However, most deep learning-based LLIE methods focus on learning the mapping function but ignore the guidance of the auxiliary priors provided by normal-light images in the training dataset, leading to unpleasant artifacts or distorted colors in the enhanced images, as shown in Figure 1.

Recently, several image restoration methods [4, 11, 38, 48] propose learning textures and details under the guidance of vector-quantized (VQ) codebook priors. VQ-based methods, e.g. VQ-VAE [24, 32] and VQ-GAN [7], usually have two stages. Specifically, in the first stage, a high-quality codebook is learned with self-reconstruction of high-quality labels and aims to record high-level semantic information and provide auxiliary priors to guide the learning of the second stage. The decoders of VQ-based networks store rich, high-quality textures and details simultaneously. However, it is unsatisfactory to employ directly VQ-based methods in LLIE. These VQ-based methods propose to construct a codebook in high-level feature space by downsampling the high-quality image with factors of 8, 16, or 32 and select one codebook item to represent high-level features, which makes image details lost and network training unstable.

To improve important details, residual quantized VAE (RQ-VAE) [16] proposes the selection of multiple codebook items to accurately represent characteristics using a residual quantization strategy. In the second stage, the low-quality images are mapped into high-level semantic space, and the codebook items closest to the high-level features are selected to be inputted to the decoder of the first stage for generating output. Taking into account the gap between low-quality and high-quality images, some VQ-based methods adopt distillation loss to allow the second-stage encoder to mimic one of the first stages. However, since these methods select the nearest codebook items relying on similarities between the feature vectors of the low-quality images and the high-quality codebook items, the quantized features are still limited and unreliable. Furthermore, the short-cut connection popularly used in the encoder-decoder structure (e.g., U-Net [31]) cannot be used in VQ-based networks, resulting in further significant detail loss.

To address these challenges, we propose a novel low-light image enhancement (LLIE) method based on VQ-VAE with a three-stage framework in this paper. In the first stage, we learn a normal-light decoder and a more hierarchical and expressive normal-light codebook by residual quantization [16]. However, constructing a more expressive codebook also means that it becomes more difficult to select the correct items from the codebook. Therefore, in the second stage, not only should the learned features from the low-light images approximate the normal-light image features by distillation loss, but also should be calculated the similarities between the low-light features and query items to select the codebook items. To avoid the loss of details caused by downsampling, we propose the third stage to protect valuable details and refine the enhanced results by a fusion branch that fuses features of the pre-trained low-light encoder and normal-light decoder on different scales. In addition, a novel brightness-aware attention module is proposed to dynamically learn the brightness and textures of images and is integrated into the fusion branch. The contributions of this paper are listed below.

- We propose a novel low-light image enhancement method based on VQ-VAE with a three-stage framework. To our knowledge, our proposed method is the first VQ-based method for LLIE.
- We construct a more hierarchical and expressive codebook by residual quantization. In addition, we design a query module to bridge the gap between low-light features and the normal-light codebook.
- To avoid the image details lost by the downsampling operation, we propose a fusion branch fusing low-light features and normal-light priors at different scales.
- We design a brightness-aware attention module that learns a brightness map to modulate features to improve the robustness of the network to the brightness.
- Extensive experimental results on several popular datasets show that our proposed method outperforms several existing state-of-the-art LLIE methods.

2. Related Work

2.1. Low-light image enhancement methods

Many conventional LLIE methods were based on histogram equalization [15, 25, 26] or gamma correction [12, 27]. These methods were aimed at expanding the dynamic range and enhancing the contrast. However, they tended to

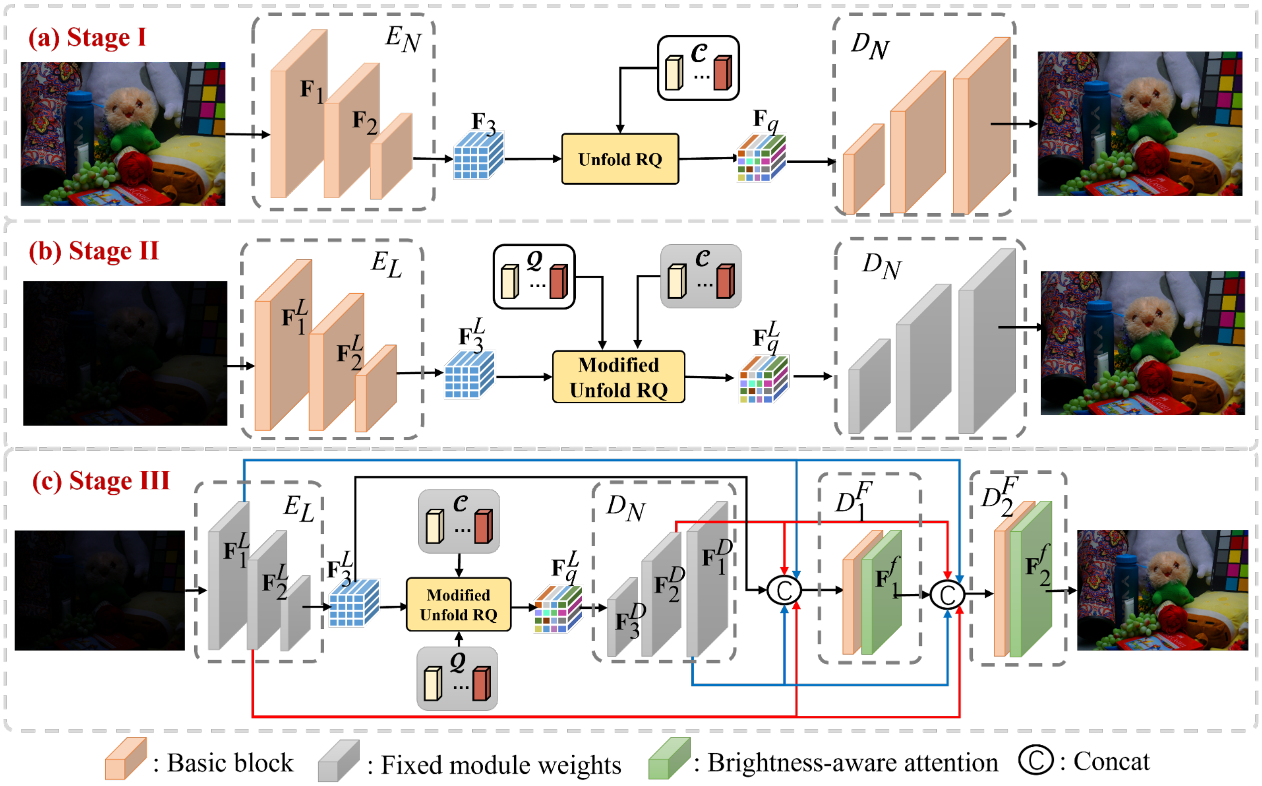


Figure 2: Architectures of the proposed three-stage framework for low-light image enhancement. (a) In stage I, we aim to learn an expressive codebook $\mathcal{C}$ and a precise normal-light decoder D_N. (b) In stage II, we learn a low-light encoder E_L and a query module $\mathcal{Q}$, while the codebook $\mathcal{C}$ and the decoder D_N are fixed. (c) In stage III, we propose a fusion branch to fuse the features of the fixed encoder E_L and decoder D_N, pursuing better results.

generate undesirable artifacts. Retinex-based methods, on the other hand, decompose an image into two components (reflectance and illumination) and then correct the illumination and suppress artifacts [14, 19, 28, 39]. Although Retinex-based methods outperformed early methods, they still show poor performance when applied to real-world images [34].

Since deep learning-based methods have gradually dominated the field of low-level image processing [5, 20, 37, 44], several works have been proposed to solve the LLIE problem [9, 17, 18, 21, 35, 42, 43, 45, 46]. Zamir et al. [45] proposed MIR-Net which integrates multiscale contextual information and preserves spatial details in high resolution. Wu et al. [40] proposed a deep unfolding network based on the traditional URetinex model. Zhang et al. [47] proposed DCC-Net to explore the color consistency between enhanced and normal-light images. Xu et al. [41] proposed a signal-to-noise (SNR)-aware network that simultaneously employs a convolution-based short-range branch and a transformer-based long-range branch for LLIE.

2.2. VQ-based image restoration methods

Previous studies have shown that better priors can lead to better restoration performance [3, 10, 22, 33]. VQ-VAE [32] introduced a codebook learned by a vector-quantized autoencoder. Since this codebook could provide a more compressed and expressive low-level feature bank, many image restoration methods propose learning image textures and details with the guidance of vector-quantized (VQ) codebook priors and achieving significant progress. Most of them obtain high-quality features by matching the vector in the codebook that is most similar to the low-quality feature. Wang et al. [38] propose using a multi-head cross-attention layer to incorporate low-quality features and high-quality priors. Gu et al. [11] propose a parallel decoder to gradually fuse the low-quality input feature and the high-quality priors. These VQ-based methods verify the effectiveness of fusing low-quality features and the corresponding high-quality priors. However, they ignore the gap between low-quality features and the codebook, which leads to inaccurate matching results. Zhou et al. [48] propose a Codeformer to predict the correct code index given the low-quality feature.

Although Codeformer improves matching accuracy, it focuses on face restoration and relies on a high compression ratio to reduce code length. Most recently, RQ-VAE [16] has faced the challenge of selecting multiple codebook elements to accurately represent features using the residual quantization strategy.

3. The Proposed Method

3.1. Overview of LLIE

To obtain normal-light images with more valuable details and less unpleasant artifacts, we propose an LLIE method based on VQ-VAE [32] with a three-stage framework, as shown in Figure 2. The components of these three stages are as follows.

- **Stage I:** The network of Stage I contains a normal-light encoder E_N, a residual quantization (RQ) module with a codebook $\mathcal{C}$, and a normal-light decoder D_N. Learning an expressive codebook $\mathcal{C}$ and a precise decoder D_N is the core of Stage I. More details will be described in Sec. 3.2.
- **Stage II:** In Stage II, we propose to learn a query module $\mathcal{Q}$ and select the codebook items according to the similarity between the features of a low-light encoder E_L and the learned query $\mathcal{Q}$. The parameters of the encoder E_L and the query $\mathcal{Q}$ require training, while the codebook $\mathcal{C}$ and the decoder D_N learned in Stage I are fixed. More details will be described in Sec. 3.3.
- **Stage III:** In Stage III, we propose a fusion branch to fuse features of the pre-trained encoder E_L and the decoder D_N. In this way, it can further protect more valuable details and obtain better performance than Stage II. More details will be described in Sec. 3.4.

As shown in Figure 2, the encoders E_L and E_N and the decoder D_N consist of three basic blocks. Each basic block, shown in Figure 3 (a), consists of a convolutional layer and three spectral attention blocks (SAB) [1, 2]. SAB [1, 2] is proposed to learn channel-wise self-attention of feature maps with low computational cost and achieve state-of-the-art spectral reconstruction performance. More details about SAB are given in the supplementary material. The encoders E_L and E_N use two downsampling operators to downsample the feature, while the decoder D_N uses two upsampling operators to recover the feature.

3.2. Stage I: Codebook $\mathcal{C}$ and Normal-light Decoder D_N Learning

In Stage I, we aim to learn more expressive normal-light image priors (i.e., the codebook $\mathcal{C} = \{\mathbf{c}_k \in \mathbb{R}^{2\times2\times c}\}_{k=1}^{K}$) by self-reconstruction of normal-light images. As shown in Figure 2 (a), the normal-light image $\mathbf{I}_N \in \mathbb{R}^{H\times W\times 3}$ is first encoded into a high-level feature $\mathbf{F}_3 \in \mathbb{R}^{h\times w\times c}$ by the encoder E_N. Then the feature $\mathbf{F}_3$ is quantized into discrete features by RQ operator [16] as shown in Figure 4 (a). We unfold the feature $\mathbf{F}_3$ into N overlapped patches $\mathcal{P} = \{\mathbf{P}_i \in \mathbb{R}^{2\times2\times c}\}_{i=1}^{N}$ and use D RQ operators to recursively discretize each patch $\mathbf{P}_i$. The d^{th} RQ operator can be described as

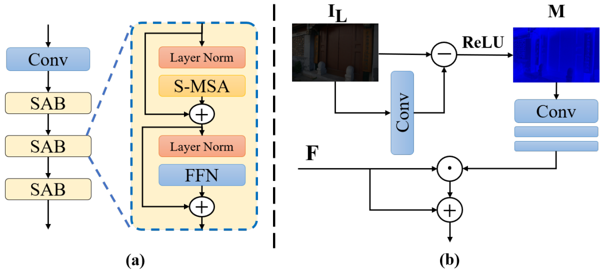


Figure 3: The architectures of (a) the basic block and (b) the proposed brightness-aware attention module.

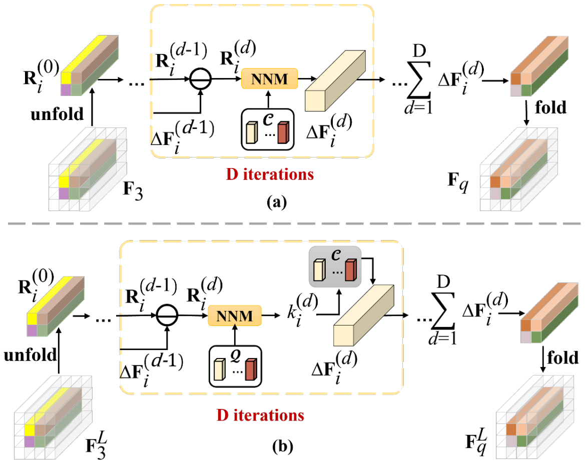


Figure 4: The procedure of (a) the Unfold RQ module and (b) the modified Unfold RQ module. NNM denotes the nearest-neighbor matching operator. Unlike the Unfold RQ module (a) use NNM operator to find the nearest codebook item $\mathbf{c}_k$, the modified Unfold RQ module (b) uses NNM operator to find the index of the nearest query item $\mathbf{q}_k$.

$$\begin{aligned} \mathbf{R}_i^{(d)} &= \mathbf{R}_i^{(d-1)} - \Delta\mathbf{F}_i^{(d-1)}, \\ \Delta\mathbf{F}_i^{(d)} &= \underset{\mathbf{c}_k\in\mathcal{C}}{\operatorname{argmin}} ||\mathbf{R}_i^{(d)} - \mathbf{c}_k||^2, \end{aligned} \tag{1}$$

where $\Delta\mathbf{F}_i^{(d)} = \mathbf{0}$, $\mathbf{R}_i^{(d)}$ denotes the d^{th} residual and $\mathbf{R}_i^{(0)} = \mathbf{P}_i$. The final discretized patch $\mathbf{P}_i^q = \sum_{d=1}^{D} \Delta\mathbf{F}_i^{(d)}$. We fold all discretized patches $\mathbf{P}_i^q$ and obtain the quantized feature $\mathbf{F}_q$ that will be fed into the normal-light decoder D_N to reconstruct the original normal-light image.

Compared to previous VQ-based face restoration methods that adopt high compression ratios (e.g., 8, 16, and 32) in the encoder, our proposed method only uses two downsampling operators to compress the input image of ratio 4 in total, greatly reducing the damage to image details and textures. Although preserving richer image details and textures means needing a larger codebook to represent them in VQ-VAE [32] and VQ-GAN [7], RQ-VAE [16] allows us to construct an expressive codebook in a small size.

Loss Function. To train the normal-light encoder E_N, the codebook $\mathcal{C}$ and the normal-light decoder D_N, we employ the total loss function $\mathcal{L} = \mathcal{L}_{recon} + \beta \cdot \mathcal{L}_{vq}$, where the reconstruction loss $\mathcal{L}_{recon}$ and the vector quantization loss $\mathcal{L}_{vq}$ are defined as:

$$\begin{aligned} \mathcal{L}_{recon} &= ||\hat{\mathbf{I}}_N - \mathbf{I}_N||_2^2, \\ \mathcal{L}_{vq} &= ||\mathrm{sg}[\mathbf{F}_3] - \mathbf{F}_q||_2^2 + \beta||\mathbf{F}_3 - \mathrm{sg}[\mathbf{F}_q]||_2^2, \end{aligned} \tag{2}$$

where β is set to 0.25 in our experiments, $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator, and $\hat{\mathbf{I}}_N$ denotes the output of the normal-light decoder D_N.

3.3. Stage II: Low-light Encoder E_L and Query $\mathcal{Q}$ Learning

In this stage, we use the low-light encoder E_L to extract the features of the low-light images and obtain the features $\mathbf{F}_1^L$, $\mathbf{F}_2^L$, and $\mathbf{F}_3^L$ on three scales, as shown in Figure 2 (b). Due to noise corruption and different brightnesses, there is a gap between low-light features $\mathbf{F}_3^L$ and normal-light codebook $\mathcal{C}$, making it difficult for low-light feature patches to match codebook items accurately. To address this issue, we introduce a query module $\mathcal{Q} = \{\mathbf{q}_k \in \mathbb{R}^{2\times2\times c}\}_{k=1}^K$ to bridge the low-light feature patches and the codebook. In other words, instead of directly calculating the distance between low-light features $\mathbf{F}_3^L$ and the normal-light codebook $\mathcal{C}$, we calculate the distance between $\mathbf{F}_3^L$ and the query $\mathcal{Q}$ and use the index of the closest distance to match the item in the codebook $\mathcal{C}$. The details of the modified RQ module are shown in Figure 4 (b) and the matching strategy can be described as

$$\begin{aligned} \mathbf{R}_i^{(d)} &= \mathbf{R}_i^{(d-1)} - \Delta\mathbf{F}_i^{(d-1)}, \\ k_i^{(d)} &= \underset{k}{\mathrm{argmin}}\, ||\mathbf{R}_i^{(d)} - \mathbf{q}_k||_2^2, \\ \Delta\mathbf{F}_i^{(d)} &= \mathbf{c}_{k_i^{(d)}}. \end{aligned} \tag{3}$$

As shown in Figure 2 (b), we input the feature $\mathbf{F}_3^L$ into the modified RQ module and obtain the quantized feature $\mathbf{F}_q^L$. Then the quantized feature $\mathbf{F}_q^L$ is fed into the pre-trained normal-light decoder D_N. The low-light encoder E_L and the query module $\mathcal{Q}$ are trainable, while the codebook $\mathcal{Q}$ and the normal-light decoder D_N are pre-trained in Stage I and fixed. The parameters of the query $\mathcal{Q}$ are initialized to be the same as the parameters of the codebook $\mathcal{C}$.

Loss Function. To train the low-light encoder E_L, we use the normal-light encoder E_H as a teacher network and E_L as a student network to distill knowledge on different scales. Specifically, we minimize the ℓ_1-distance between the feature maps $\mathbf{F}_j$ and $\mathbf{F}_j^L$. In stage II, we use the following loss function:

$$\mathcal{L} = \mathcal{L}_{distill} + \mathcal{L}_{query}, \tag{4}$$

$$\mathcal{L}_{distill} = \sum_{j=1}^{3} ||\mathbf{F}_j^L - \mathbf{F}_j||_1, \tag{5}$$

$$\mathcal{L}_{query} = ||\mathrm{dis}(\mathbf{F}_3, \mathcal{C}) - \mathrm{dis}(\mathrm{sg}[\mathbf{F}_3^L], \mathcal{Q})||_1, \tag{6}$$

where $\mathrm{dis}(\mathbf{F}_3, \mathcal{C})$ and $\mathrm{dis}(\mathrm{sg}[\mathbf{F}_3^L], \mathcal{Q})$ denote the distance maps between the unfolded patches $\mathcal{P}$ of $\mathbf{F}_3$ / $\mathbf{F}_3^L$ and the items of $\mathcal{C}$ / $\mathcal{Q}$, respectively. Under the constraint of $\mathcal{L}_{query}$, the distance map of $\mathbf{F}_3^L$ and $\mathcal{Q}$ will be close to the distance map of $\mathbf{F}_3$ and $\mathcal{C}$.

3.4. Stage III: Feature Fusion

In Stage II, we have constructed the normal-light images preliminarily. However, due to the lack of a shortcut connection between the encoder E_L and the decoder D_N, it is inevitable that many valuable details and textures will be lost by the downsampling operator.

To compensate for the loss of detail and textures, we introduce a fusion branch that fuses features from the encoder E_L and the decoder D_N on different scales to generate high-quality normal-light images. In addition, considering that the brightness of the image may vary in different areas, we propose a brightness-aware attention module in the fusion branch, making the network robust to brightness.

Fusion Branch. We propose fusion of features of the encoder E_L and the decoder D_N on different scales by two fusion blocks D_1^F and D_2^F. Each fusion block consists of a basic block and a brightness-aware attention module. As shown in Figure 2 (c), the fusion branch can be described as

$$\begin{aligned} \mathbf{F}_1^{cat} &= \mathrm{Concat}(\mathbf{F}_3^L \uparrow, \mathbf{F}_2^L, \mathbf{F}_1^L \downarrow, \mathbf{F}_2^D, \mathbf{F}_1^D \downarrow), \\ \mathbf{F}_1^f &= \mathrm{BA}_1(D_1^F(\mathbf{F}_1^{cat}), \mathbf{M} \downarrow), \\ \mathbf{F}_2^{cat} &= \mathrm{Concat}(\mathbf{F}_1^f \uparrow, \mathbf{F}_2^L \uparrow, \mathbf{F}_1^L, \mathbf{F}_2^D \uparrow, \mathbf{F}_1^D), \\ \mathbf{F}_2^f &= \mathrm{BA}_2(D_2^F(\mathbf{F}_2^{cat}), \mathbf{M}), \\ \mathbf{I}_{rec} &= \mathrm{Conv}(\mathbf{F}_2^f), \end{aligned} \tag{7}$$

where ↑ and ↓ denote the upsampling and downsampling operators with factor 2 respectively, $\mathrm{Concat}(\cdot, \cdot, ...)$ concatenates all input features, Conv denotes a 3×3 Conv layer to reconstruct the final normal-light images $\mathbf{I}_{rec}$, $\mathbf{M}$ denotes a brightness map, and $\mathrm{BA}_1(\cdot,\cdot)$ and $\mathrm{BA}_2(\cdot,\cdot)$ denote the brightness-aware attention module.

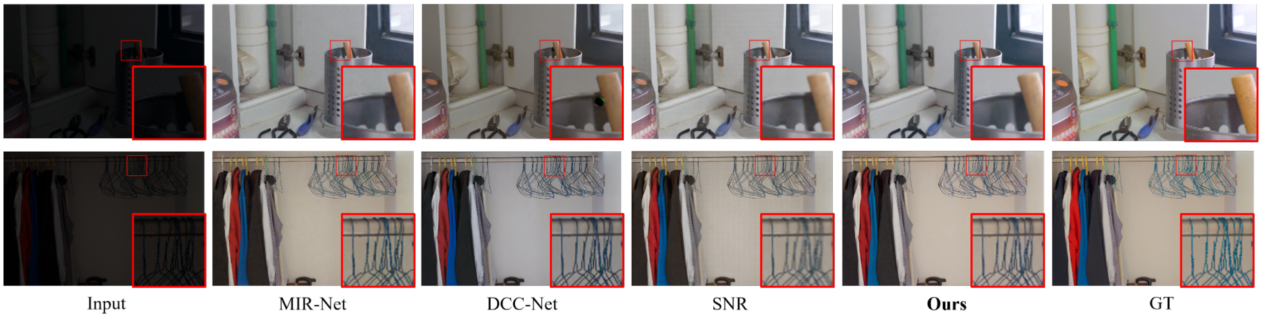


Figure 5: Visual quality comparisons of different low-light image enhancement methods on the LOLv1 dataset.

Brightness-aware Attention Module. To make the network robust to brightness, we propose a new brightness-aware attention module, as shown in Figure 3 (b). First, we calculate the brightness map as follows.

$$\mathbf{M} = \mathrm{ReLU}(\mathbf{I}_L - \mathrm{Conv}(\mathbf{I}_L)). \tag{8}$$

The brightness map **M** marks different brightness levels and is utilized to generate spatial attention to modulate fused features such as

$$\mathrm{BA_j}(\mathbf{F}, \mathbf{M}) = \mathbf{F} + \mathbf{F} \odot \mathrm{Convs}(\mathbf{M}), \tag{9}$$

where $\mathrm{Convs}(\cdot)$ denotes three 3×3 Conv layers, **F** denotes the fused feature, and $j = 1$ or 2.

Loss Function. In Stages III, the parameters of the encoder E_L and the decoder D_N are fixed, while the fusion branch parameters will be optimized by minimizing the ℓ_1 loss as

$$\mathcal{L} = ||\mathbf{I}_{rec} - \mathbf{I}_N||_1. \tag{10}$$

4. Experiments

4.1. Datasets, metrics and implementation details

To verify the performance of our proposed low-light image enhancement method, we conducted extensive experiments on three public datasets (i.e., the LOLv1 [39] dataset, the LOLv2-Real [43] dataset, and the LOLv2-Synthetic [43] dataset). The LOLv1 [39] dataset has 500 low/normal-light image pairs and is divided into 485 pairs for training and 15 pairs for testing. The LOLv2-Real [43] and LOLv2-Synthetic [43] datasets are larger and more diverse, both including 689 low/normal-light image training pairs and 100 test pairs. Two commonly used metrics, i.e., peak signal-to-noise (PSNR) and structural similarity (SSIM) [36], are adopted to evaluate the performance of competing low-light image enhancement methods.

We implement our method in PyTorch[23] with 2 NVIDIA 3090 GPUs. We adopt Adam [13] optimizer and standard data augmentation (e.g., vertical and horizontal flips) for the whole training procedure. We set the codebook size K to 1024 and each codebook item has a size of $2 \times 2 \times 256$. We implement the unfold RQ module and its modified counterpart with 8 and 6 residual quantization operators on the LOLv1 dataset and the LOLv2-Real and LOLv2-Synthetic datasets, respectively.

4.2. Compared with state-of-the-art methods

We compare the proposed method with eleven state-of-the-art deep learning-based methods, including LPNet [18], MIR-Net [45], A3DLUT [35], Band [42], Retinex [21], Sparse [43], IPT [5], Uformer [37], SNR [41], URetinex-Net [40] and DCC-Net [47]. Note that our experimental setup is consistent with SNR [41] and the numerical results of competing methods are cited from SNR [41]. Since the source code of URetinex-Net [40] and DCC-Net [47] are not public, we have not compared the proposed method with them on the LOLv2-Real and LOLv2-Synthetic datasets.

Quantitative Analysis. The numerical results of all competing LLIE methods on the LOLv1, LOLv2-Real, and LOLv2-Synthetic datasets are shown in Tables 1 and 2. From Tables 1 and 2, we can observe that our proposed method achieves the best average PSNR and SSIM results. On the LOLv1, LOLv2-Real, and LOLv2-Synthetic datasets, our proposed method outperforms the second-best method SNR [41] by 0.63 dB, 0.89 dB, and 1.8 dB on average PSNR, respectively. The improvements by our methods over Band [42], Sparse [43] and MIR-Net [45] are (5.11 dB, 2.08 dB, and 2.72 dB), (8.04 dB, 2.31 dB, and 3.89 dB), and (1.1 dB, 2.35 dB, and 4 dB) on average, respectively.

Visual Analysis. Figures 5-7 show the visual comparison of several LLIE methods on the LOLv1, LOLv2-Real and LOLv2-Synthetic datasets. MIR-Net [45] produces blurred artifacts as shown in Figures 6 and 7 (Bottom). The color tone of the DCC-Net [47] and SNR [41] results is inconsistent with the ground truth, as shown in Figures 5 and 7 (Top), respectively. Our proposed method can generate more pleasant images with more details and textures of the image, fewer undesirable blurred artifacts, and better color

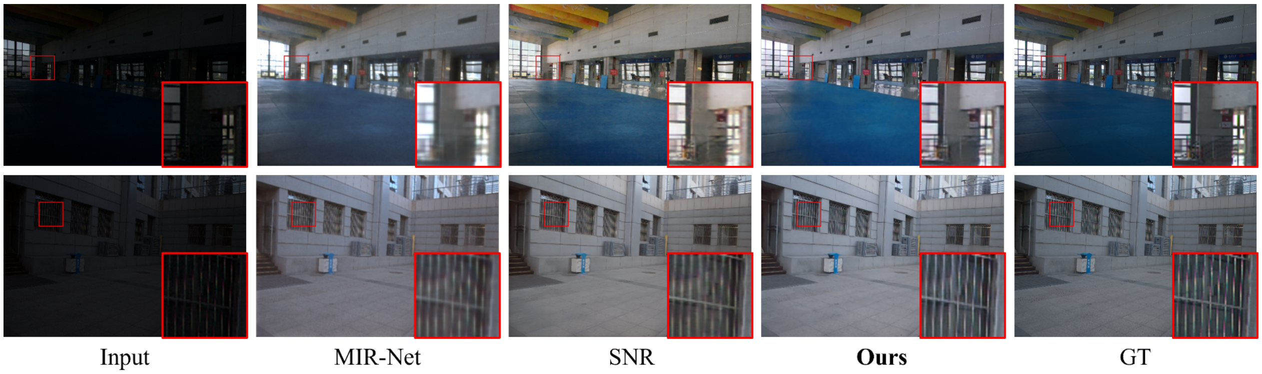


Figure 6: Visual quality comparisons of different low-light image enhancement methods on the LOLv2-Real dataset.

Methods	PSNR	SSIM
A3DLUT [35]	14.77	0.458
IPT [5]	16.27	0.504
Uformer [37]	16.36	0.507
Sparse [43]	17.20	0.640
Retinex [21]	18.23	0.720
Band [42]	20.13	0.830
URetinex-Net [40]	21.33	0.835
LPNet [18]	21.46	0.802
DCC-Net [47]	22.72	0.810
MIR-Net [45]	24.14	0.830
SNR [41]	24.61	0.842
ours	**25.24**	**0.855**

Table 1: Quantitative comparison on the LOLv1 dataset.

	LOLv2-Real		LOLv2-Syn	
Methods	PSNR	SSIM	PSNR	SSIM
LPNet [18]	17.80	0.792	19.51	0.846
Retinex [21]	18.37	0.723	16.55	0.652
A3DLUT [35]	18.19	0.745	18.92	0.838
Uformer [37]	18.82	0.771	19.66	0.871
IPT [5]	19.80	0.813	18.30	0.811
MIR-Net [45]	20.02	0.820	21.94	0.876
Sparse [43]	20.06	0.816	22.05	0.905
Band [42]	20.29	0.831	23.22	0.927
SNR [41]	21.48	0.849	24.14	0.928
ours	**22.37**	**0.854**	**25.94**	**0.941**

Table 2: Quantitative comparison on the LOLv2-Real and LOLv2-Synthetic dataset.

consistency. More comparisons are shown in the supplementary material.

4.3. Ablation Study

To verify the impacts of residual quantization (RQ), query module $\mathcal{Q}$, fusion branch (FB), and brightness-aware attention (BA) module, we conduct several ablation studies on the LOLv2-Real dataset. We set the VQ-VAE method as the baseline, which does not use FB, $\mathcal{Q}$, and BA, and construct the codebook by vector quantization. The results of the ablation studies are shown in Table 3.

RQ-based Codebook. Compared model A with the baseline, constructing the codebook by residual quantization can improve the PSNR and SSIM results by 0.37 dB and 0.055. From Figure 8, we can observe that model A result has more image textures than the baseline. This proves that using residual quantization can learn a more expressive codebook.

Query Module $\mathcal{Q}$. From Table 3, we can observe that the improvement of model B over model A is 0.41 dB. The visual result of model B has fewer blurry artifacts than that of model A, as shown in Figure 8.

Fusion Branch. Since the details and textures of the image are lost in the encoder, we propose the fusion branch to preserve the textures and details of the image that can be verified from Figure 8. As shown in Table 3, compared to model C with model B, the fusion branch without the brightness-aware attention module brings an improvement of 0.15 dB.

Brightness-aware Attention module. The brightness-aware attention module further improves the PSNR results by 0.38 dB. We also show the learned brightness map **M** in Figure 8. The brightness map **M** is consistent with the texture and brightness of the image. As shown in Figure 8, the visual result of model D is more consistent with the GT in color tone, such as the areas on the wall and the table.

Models	RQ	Query	FB	BA	PSNR	SSIM
Baseline					21.06	0.747
A	✓				21.43	0.802
B	✓	✓			21.84	0.794
C	✓	✓	✓		21.99	0.853
D	✓	✓	✓	✓	**22.37**	**0.854**

Table 3: Results of the ablation studies.

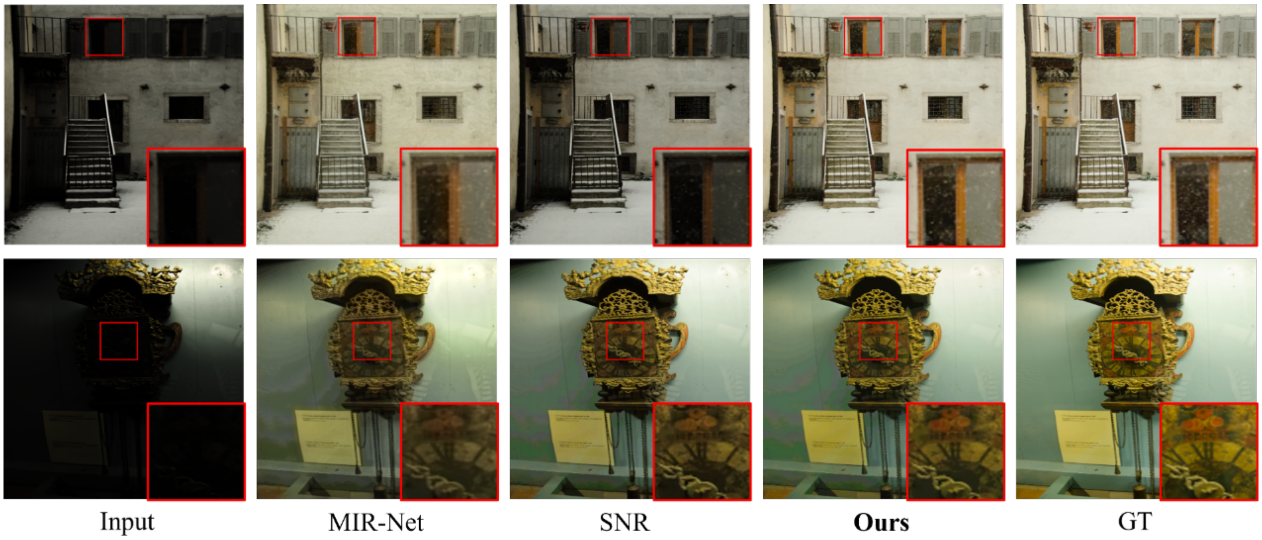


Figure 7: Visual quality comparisons of different low-light image enhancement methods on the LOLv2-Synthetic dataset.

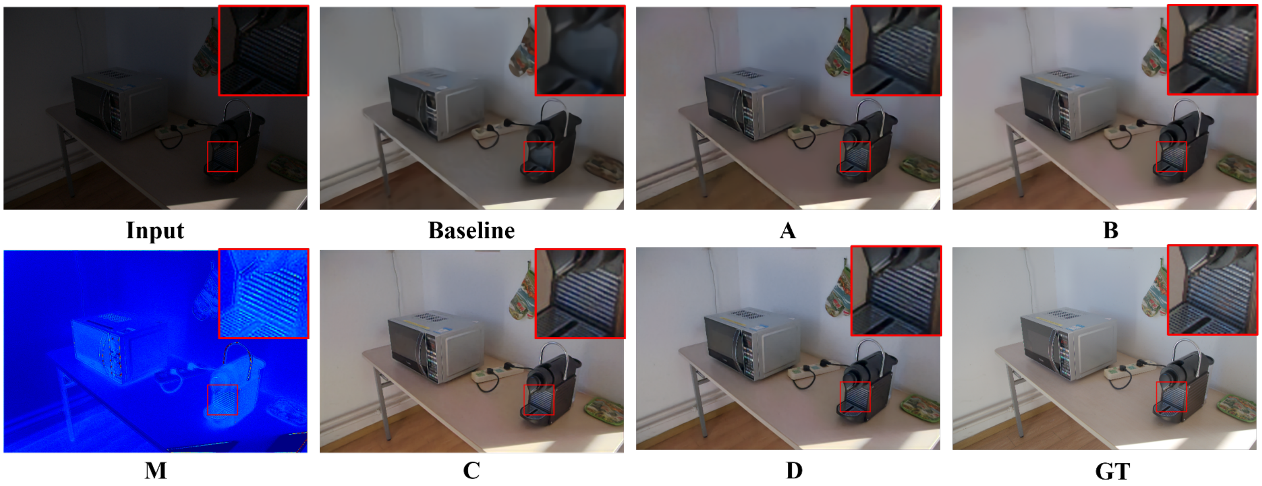


Figure 8: Visual comparisons of the ablation studies.

5. Limitations and Conclusion

In this paper, we propose a novel low-light image enhancement method based on VQ-VAE with a three-stage framework. Unlike VQ-VAE, we propose learning a normal-light codebook by residual quantization. Due to the gap between the low-light features and the codebook items, we have developed a query module to address this gap. To preserve the textures and details of the image, we propose a branch of fusion to fuse the encoder and decoder features. Further improvement of color consistency is achieved by a brightness-aware attention module integrated into the fusion branch. Extensive experiments on real-captured and synthetic datasets demonstrate that our proposed method outperforms existing state-of-the-art low-light image enhancement methods. Although our proposed method achieves promising results, there is still room for improvement. In this paper, the basic block including three spectral-wise attention blocks only learns channel-wise self-attention and ignores spatial self-attention. In the future, we will integrate the spatial self-attention module into the basic block for producing high-quality images. Furthermore, iterative optimization of stage II and stage III in combination may be a good way to further improve results.

Acknowledgments. This work was supported in part by the Natural Science Foundation of China under Grant 61991451 and Grant 61836008. Xin Li's work is partially supported by the NSF under grants IIS-2114664, CMMI-2146076, and CCSS-2318758.

References

[1] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17502–17511, 2022.

[2] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 745–755, 2022.

[3] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021.

[4] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1329–1338, 2022.

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[9] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2782–2790, 2016.

[10] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020.

[11] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 126–143. Springer, 2022.

[12] Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE transactions on image processing*, 22(3):1032–1041, 2012.

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

[14] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.

[15] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE transactions on image processing*, 22(12):5372–5384, 2013.

[16] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[17] Chongyi Li, Jichang Guo, Fatih Porikli, and Yanwei Pang. Lightennet: A convolutional neural network for weakly illuminated image enhancement. *Pattern recognition letters*, 104:15–22, 2018.

[18] Jiaqian Li, Juncheng Li, Faming Fang, Fang Li, and Guixu Zhang. Luminance-aware pyramid network for low-light image enhancement. *IEEE Transactions on Multimedia*, 23:3153–3165, 2020.

[19] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018.

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10561–10570, 2021.

[22] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020.

[23] 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. *Advances in neural information processing systems*, 32, 2019.

[24] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021.

[25] Stephen M Pizer. Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene

johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group. In *Proceedings of the first conference on visualization in biomedical computing, Atlanta, Georgia*, volume 337, page 1, 1990.

[26] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[27] Shanto Rahman, Md Mostafijur Rahman, Mohammad Abdullah-Al-Wadud, Golam Dastegir Al-Quaderi, and Mohammad Shoyaib. An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, 2016(1):1–13, 2016.

[28] Zia-ur Rahman, Daniel J Jobson, and Glenn A Woodell. Retinex processing for automatic image enhancement. *Journal of Electronic imaging*, 13(1):100–110, 2004.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[32] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[33] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2747–2757, 2020.

[34] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6849–6857, 2019.

[35] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2471–2480, 2021.

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[37] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022.

[38] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022.

[39] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2022.

[41] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17714–17724, 2022.

[42] 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 Transactions on Image Processing*, 30:3461–3473, 2021.

[43] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021.

[44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[45] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020.

[46] 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 Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2058–2073, 2020.

[47] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. Deep color consistent network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1899–1908, 2022.

[48] Shangchen Zhou, Kelvin CK Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *arXiv preprint arXiv:2206.11253*, 2022.