VQCNIR: Clearer Night Image Restoration with Vector-Quantized Codebook

Wenbin Zou[1], Hongxia Gao[1,2*], Tian Ye[3], Liang Chen[4], Weipeng Yang[1], Shasha Huang[1], Hongshen Chen[1], Sixiang Chen[3]

[1]The School of Automation Science and Engineering, South China University of Technology, Guangzhou
[2]Research Center for Brain-Computer Interface, Pazhou Laboratory, Guangzhou
[3]The Hong Kong University of Science and Technology, Guangzhou
[4]College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou
hxgao@scut.edu.cn

Abstract

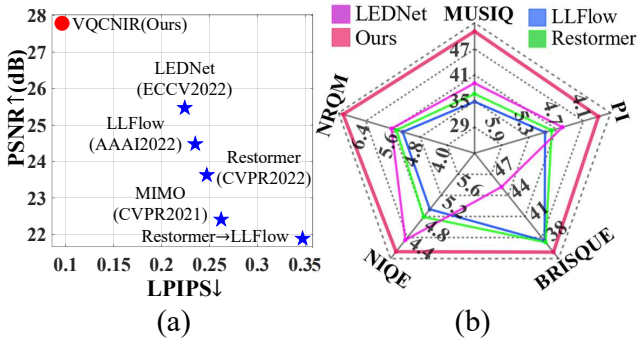
Night photography often struggles with challenges like low light and blurring, stemming from dark environments and prolonged exposures. Current methods either disregard priors and directly fitting end-to-end networks, leading to inconsistent illumination, or rely on unreliable handcrafted priors to constrain the network, thereby bringing the greater error to the final result. We believe in the strength of data-driven high-quality priors and strive to offer a reliable and consistent prior, circumventing the restrictions of manual priors.

In this paper, we propose Clearer Night Image Restoration with Vector-Quantized Codebook (VQCNIR) to achieve remarkable and consistent restoration outcomes on real-world and synthetic benchmarks. To ensure the faithful restoration of details and illumination, we propose the incorporation of two essential modules: the Adaptive Illumination Enhancement Module (AIEM) and the Deformable Bi-directional Cross-Attention (DBCA) module. The AIEM leverages the inter-channel correlation of features to dynamically maintain illumination consistency between degraded features and high-quality codebook features. Meanwhile, the DBCA module effectively integrates texture and structural information through bi-directional cross-attention and deformable convolution, resulting in enhanced fine-grained detail and structural fidelity across parallel decoders.

Extensive experiments validate the remarkable benefits of VQCNIR in enhancing image quality under low-light conditions, showcasing its state-of-the-art performance on both synthetic and real-world datasets. The code is available at https://github.com/AlexZou14/VQCNIR.

Introduction

To obtain reliable images in night scenes, long exposure is often used to allow more available light to illuminate the image. However, images captured in this way still suffer from low visibility and color distortion issues. Moreover, long exposure is susceptible to external scene disturbances, such as camera shake and dynamic scenes, which can cause motion blur and noise in the images (2022). Therefore, night images often exhibit complex degradation problems (2022a; 2022; 2023) such as low illumination and blur, making the recovery of high-quality images with realistic texture and normal lighting conditions extremely challenging.



(a) (b)

Figure 1: Quantitative comparisons with state-of-the-art methods. (a) PSNR and LPIPS results on the LOL-Blur dataset. (b) Results for five perceptual metrics on the Real-LOL-Blur dataset. For PSNR, MUSIQ (2021), and NRQM (2017) higher is better, while lower is better for LPIPS (2018a), NIQE (2012), BRISQUE (2012), and PI (2018).

With the great success of deep learning methods (2022; 2023d; 2023b; 2023c; 2023; 2023a) in image restoration, numerous deep learning-based algorithms have been proposed to tackle this challenging task. Currently, most researchers only consider the low-light problem in night images and have proposed numerous low-light image enhancement (LLIE) methods (2017; 2018; 2019; 2020; 2021; 2019; 2021). Although these LLIE methods can produce visually pleasing results, their generalization ability is limited in real night scenes. This is mainly attributed to the fact that LLIE methods focus primarily on enhancing image luminance and reducing noise while ignoring the spatial degradation caused by blur that leads to ineffective recovery of sharp images. An intuitive idea is to combine image deblurring methods with LLIE methods to address this problem. However, most existing deblurring methods (2021; 2021; 2019; 2022; 2022b) are trained on datasets captured under normal illumination conditions, which makes them not suitable for night image deblurring. In particular, due to the poor visibility in dark regions of night images, these methods may fail to effectively capture motion blur cues, resulting in unsatisfactory deblurring performance. Therefore, simply cascading

*Corresponding author.

LLIE and deblurring methods do not produce satisfactory recovery results. To better handle the joint degradation process of low illumination and blur, Zhou et al. (2022) first proposed a LOL-Blur dataset and an end-to-end encoder-decoder network called LEDNet. LEDNet can achieve high performance on the synthetic LOL-Blur dataset. However, its generalization ability in real scenes is still limited.

The aforementioned night restoration methods have difficulties in recovering correct textures and reliable illuminations from low-quality night images. ***This is due to the lack of stable and reliable priors, as most existing priors are generated from low-quality images.*** For instance, Retinex-based techniques (2018; 2019; 2021) employ illumination estimation through the decomposition of low-quality images, while blur kernels are estimated using the same degraded inputs. However, the biased estimation of priors leads to cumulative errors in the final outcomes.

Therefore, we introduce the vector quantization (VQ) codebook as a credible and reliable external feature library to provide high-quality priors for purely data-driven image restoration, instead of relying on vulnerable handcrafted priors.

The VQ codebook is an implicit prior generated by a VQGAN (2021) and trained on a vast corpus of high-fidelity clean images. Hence, a well-trained VQ codebook can provide comprehensive, high-quality priors for complex degraded images, effectively addressing complex degradation. ***Furthermore, inconsistent illumination and incorrect matching between the degradation features of night images and the pristine features in the VQ codebook can lead to unsatisfactory visual effects when directly reconstructing using the codebook.*** It may even amplify blur and produce artifacts in restored images. Hence, the pivotal step towards harnessing codebook priors for the restoration of night-blurred images lies in precisely aligning the high-quality codebook features.

In this paper, we propose a novel method called **Clearer Night Image Restoration with Vector-Quantized Codebook (VQCNIR)** for night image restoration. To address the aforementioned key considerations, our proposed VQCNIR incorporates two purpose-built modules. Specifically, we design the **Adaptive Illumination Enhancement Module (AIEM)** that leverages inter-channel correlations of features to estimate curve parameters and adaptively enhances illumination in the features. This effectively addresses inconsistent illumination between degraded features and high-quality VQ codebook features. To ameliorate feature mismatch between degraded and high-quality features, we propose a parallel decoder integrating **Deformable Bi-directional Cross-Attention (DBCA)**. This parallel design effectively incorporates high-quality codebook features while efficiently fusing texture and structural information from the parallel encoder. Our proposed DBCA performs context modeling between high and low-quality features, adaptively fusing them to gradually recover fine details that enhance overall quality. As depicted in Figure 1, our method not only achieves superior performance on synthetic data, but also generalizes well to real-world scenes. Extensive experiments on publicly available datasets demonstrate that our method surpasses existing state-of-the-art methods on both distortion and perceptual metrics.

Our key contributions are summarized as follows:

- We propose VQCNIR, a new framework that formulates night image restoration as a matching and fusion problem between degraded and high-quality features by introducing a high-quality codebook prior. This addresses limitations of previous methods that rely solely on low-quality inputs, and achieves superior performance.
- We propose an adaptive illumination enhancement module that utilizes the inter-channel dependency to estimate curve parameters. This effectively addresses the inconsistency of illumination between the degraded features and high-quality VQ codebook features.
- We further propose a deformable bi-directional cross-attention, which utilizes a bi-directional cross-attention mechanism and deformable convolution to address the misalignment issue between features from the parallel decoder and restore the more accurate texture details.

Related Work

Image Deblurring

Recent advances in deep learning techniques have greatly impacted the field of computer vision. A large number of deep learning methods have been proposed for both single image and video deblurring tasks (2014; 2017; 2018; 2019; 2019; 2020; 2021), and have demonstrated superior performance. With the introduction of large training datasets for deblurring tasks (2009; 2017; 2019), many researchers (2009; 2019) have adopted end-to-end networks to directly recover clear images. Despite the fact that end-to-end methods outperform traditional approaches, they may not be effective in cases with severe blurring. To improve network performance, some methods (2017; 2018; 2021) use multi-scale architectures to enhance deblurring at different scales.

However, the limited ability of these methods to capture the correct blur cues in low-light conditions, particularly in dark areas, has hindered their effectiveness in handling low-light blurred images. To tackle this issue, Zhou et al. (2022) introduce a night image blurring dataset and develop an end-to-end UNet architecture that incorporates a learnable non-linear layer to effectively enhance dark regions without over-exposing other areas.

Low-Light Image Enhancement

Recent years have witnessed the impressive success of deep learning-based low-light image enhancement (LLIE) since the first pioneering work (2022). Many end-to-end methods (2017; 2019) have been proposed for enhancing image illumination using an encoder-decoder framework. To further improve the performance of LLIEs, researchers have developed deep Retinex-based methods (2018; 2019; 2021) inspired by Retinex theory, which employs dedicated subnetworks to enhance the illuminance and reflectance components and achieve better recovery performance. However, such methods have limitations, as the enhancement results strongly depend on the characteristics of the training data.

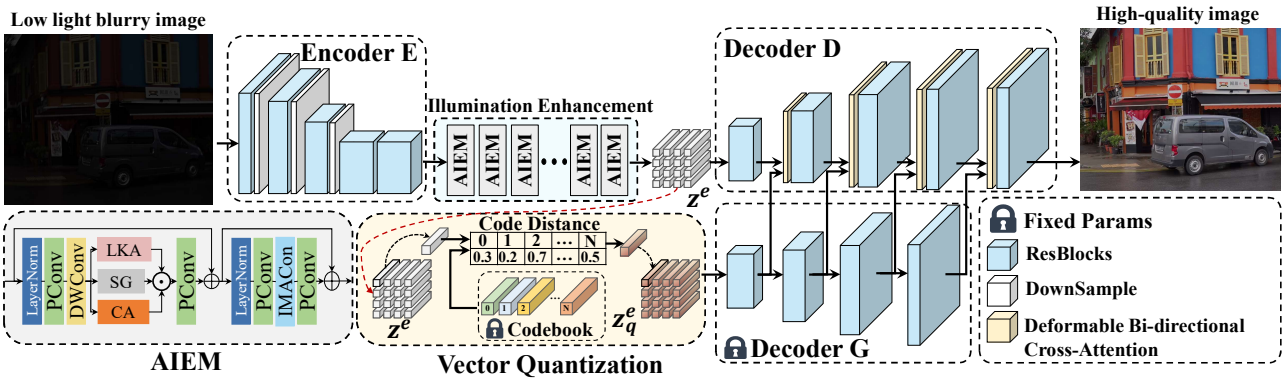


Figure 2: The framework of the proposed VQCNIR. It consists of an encoder, some adaptive illumination enhancement modules (AIEM), and a parallel decoder with a deformable bi-directional cross-attention (DBCA), allowing the network effectively to exploit high-quality codebook prior information.

To improve the generalization ability of the network, researchers (2020; 2021; 2021) propose a number of unsupervised methods. For example, Jiang et al (2021) introduced self-regularization and unpaired training into LLIE with EnlightenGAN. Additionally, Guo et al (2020) propose a fast and flexible method for estimating image enhancement depth curves that do not require any normal illumination reference images during the training process.

Verctor-Quantized Codebook

VQVAE (2017) is the first to introduce vector quantization (VQ) techniques into an autoencoder-based generative model to achieve superior image generation results. Specifically, the encoded latent variables are quantized to their nearest neighbors in a learnable codebook, and the resulting quantized latent variables are used to reconstruct the data samples. Building upon VQVAE, subsequent work has proposed various improvements to codebook learning. For instance, VQGAN (2021) utilizes generative adversarial learning and refined codebook learning to further enhance the perceptual quality of reconstructed images. The well-trained codebook can serve as a high-quality prior that can be leveraged for various image restoration tasks such as image super-resolution and face restoration. To this end, Chen et al. (2022a) introduce a VQ codebook prior for blind image super-resolution, which matches distorted LR image features with distortion-free HR features from a pre-trained HR prior. Furthermore, Gu et al. (2022) explore the impact of internal codebook properties on reconstruction performance and extended discrete codebook techniques to face image restoration. Drawing inspiration from these works, we apply the high-quality codebook prior to night image restoration.

Methodology

Framework Overview

To improve the recovery of high-quality images with realistic textures and normal illumination from night image x containing complex degradation, we introduce a Vector-Quantized codebook as high-quality prior information to design a night image restoration network (VQCNIR). The overview of the VQCNIR framework is illustrated in Figure 2. VQCNIR comprises an encoder E, an adaptive illumination enhancement module, a high-quality codebook $\mathcal{Z}$, and two decoders G and D. Decoder G is a pre-trained decoder from VQGAN with fixed parameters. Decoder D represents the primary decoder, which progressively recovers fine details by fusing high-quality features in decoder G.

VQ Codebook for Priors

VQ Codebook: We first briefly describe the VQGAN (2021) model and its codebook, and more details can be referenced in (2021). Given a high-quality image $x_h \in \mathbb{R}^{H\times W\times 3}$ with normal light, the encoder E maps the image x_h to its spatial latent representation $\hat{z} = E(x) \in \mathbb{R}^{h\times w\times n_z}$, where n_z is the dimension of latent vectors. Then, each element $\hat{z}_i \in \mathbb{R}^{n_z}$ Euclidean distance nearest vector z_k in the codebook is found as a VQ representation $z_{\mathbf{q}}$ by the element-by-element quantization process $\mathbf{q}(\cdot)$. It is shown as follows:

$$z_{\mathbf{q}} = \mathbf{q}(\hat{z}) := \left(\arg\min_{z_k\in\mathcal{Z}} ||\hat{z}_i - z_k||_2^2 \right) \in \mathbb{R}^{h\times w\times n_z}, \quad (1)$$

where the codebook is $\mathcal{Z} = \{z_k\}_{k=1}^{K} \in \mathbb{R}^{K\times n_z}$ with K discrete codes. Then, the decoder G maps the quantized representation $z_{\mathbf{q}}$ back into sRGB space. The overall reconstruction process can be formulated as follows:

$$\hat{x}_h = G(z_{\mathbf{q}}) = G(\mathbf{q}(E(x))) \approx x_h, \quad (2)$$

VQ Codebook for Night Image Restoration: To fully explore the effect of the VQ codebook prior on night image restoration, several preliminary experiments were implemented to analyze the advantages and disadvantages of VQGAN. First, we use the well-trained VQGAN to reconstruct the real image. The experimental results are shown in Figure 3 top. From the figure, we can see that VQGAN can generate vivid texture details in the reconstructed images. However, some of the structural information is lost in the vector quantization process, resulting in distortion and artifacts in the reconstructed image. Therefore, the reconstruction solely depends on the quantized features in the

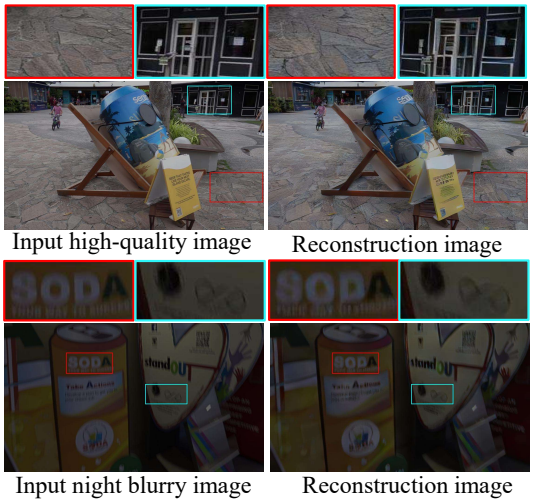


Figure 3: VQGAN reconstruction results. On the left is the input image and on the right is the reconstructed image. VQGAN can provide rich detail for high-quality images but can cause some structural distortion. In degraded images, image distortion is worsened because the degraded features do not match the correct high-quality codebook features.

codebook and does not yield satisfactory recovery results. The most intuitive idea is to combine the texture information generated by the quantized features from the codebook with the structural information of the latent representation to avoid structural distortion of the image.

Subsequently, we explore the effectiveness of VQGAN trained on high-quality images for the reconstruction of degraded images at night. As shown in Figure 3 bottom, the restoration image is unable to recover to normal illumination due to the inconsistent illumination of the input image and the training set of VQGAN. Moreover, we found that VQGAN further deteriorates blurred textures and produces artifacts. This was attributed to the difficulty of the network in matching the correct VQ codebook features, which resulted in the inability of VQGAN under high-quality image training to recover from low illumination and blur. Therefore, we design an adaptive illumination enhancement module and a deformable bi-directional cross-attention for the mentioned low light and blur problems respectively.

Adaptive Illumination Enhancement

Based on the previous observations and analysis, we design an Adaptive Illumination Enhancement Module (AIEM) to solve the problem of illumination inconsistency between the quantized features and the latent features obtained from the encoder, as shown in Figure 2. This module consists of two parts: Hierarchical Information Extraction (HIE) and Illumination Mutual Attention Enhancement (IMAE).

Hierarchical Information Extraction: Local lighting, such as light sources, is often observed in night-time environments. However global operation often over- or under-enhances these local regions. Thus, we employ channel attention and large kernel convolution attention to extract spatial information at different hierarchies. Specifically, HIE first employs layer normalization to stabilize the training and then performs spatial information fusion of different receptive fields. A residual shortcut is used to facilitate training convergence. Following the normalization layer, the point-wise convolution and the 3×3 depth-wise convolution are used to capture spatially invariant features. Then, three parallel operators are used to aggregate channel and spatial information. The first operator uses SimpleGate (2022b) to apply non-linear activation on spatially invariant features. The second operator is channel attention (2018b) to modulate the feature channels. The third one is the large kernel convolution attention (2022) to handle spatial features. The three branches output feature maps of the same size. Point-wise multiplication is used to fuse the diverse feature from the three branches directly. Finally, the output features are adjusted by point-wise convolution.

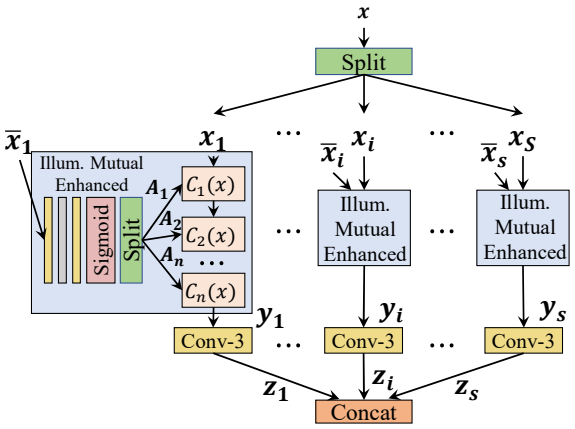


Figure 4: The architecture of Illumination Mutual Attention Convolution (IMAConv).

Illumination Mutual Attention Enhancement: According to the hierarchical information of different receptive fields obtained from the HIE, IMAE first utilize layer normalization to stabilize the training, and then illumination enhancement was applied to the features. Specifically, we design the novel illumination mutual attention convolution (IMAConv) that uses the dependencies between feature channels to estimate the curve parameters and thus adjust the illumination of the features. Two point-wise convolutions are used to adjust the input and output features of IMAConv. Residual connections are used to facilitate training convergence.

Illumination Mutual Attention Convolution: Considering that the illumination variation is similar between feature channels, we inspired by Zero-DCE (2020) introduce curve estimation and channel mutual mapping to propose an illumination mutual attention convolution that adjusts the pixel range of the feature to enhance the illumination, as shown in Figure 4. Specifically, given the input features of IMAConv as $x_f \in \mathbb{R}^{C_{in} \times H_f \times W_f}$. we first divide x into S parts at a time along the channel as follows:

$$x_f^1, x_f^2, ..., x_f^S = \text{split}(x_f), \tag{3}$$

where split$(\cdot)$ denotes the split operation. For each part $x_f^i \in$

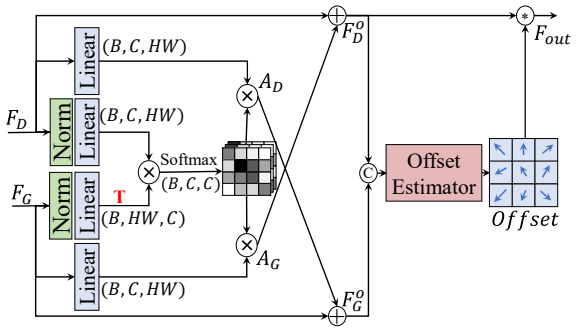


Figure 5: The architecture of deformable bi-directional cross-attention (DBCA). The offset estimator module consists of a number of large kernel convolutions that use information from the large receptive fields to help fuse the features of the two decoders.

$\mathbb{R}^{\frac{C_{in}}{S} \times H_f \times W_f}$, we concatenate the channel features excluding x_f^i together as the complimentary to x_i, denoted as $\bar{x}_f^i$. Both x_f^i and $\bar{x}_f^i$ are passed into the illumination mutual enhanced, which estimates multiple curve parameters $A = \{A_i\}_{i=1}^N$ through the curve estimation network $\mathcal{F}$. A_i is used to adjust the range of pixel values from the features. The whole process is formulated as:

$$A_1, A_2, ..., A_n = \text{split}(\mathcal{F}(\bar{x}_i)), \tag{4}$$

$$y_f^i = C_n(x_f^i, A_1, A_2, \cdots A_n), \tag{5}$$

where $\mathcal{F}(\cdot)$ and $C_n(\cdot)$ denote the curve estimation network and the high-order curve mapping function. The curve estimation network $\mathcal{F}$ consists of three convolutional layers with kernel sizes of 5, 3, and 1, respectively, two activation functions, and a sigmoid function. For the high-order curve mapping function C_n, it can be formulated as follows:

$$C_n(x_f^i) = \begin{cases} A_1 x_f^i(1 - x_f^i) + x_f^i, n = 1 \\ A_{n-1} C_{n-1}(x_f^i)(1 - C_{n-1}(x_f^i)) + C_{n-1}(x_f^i), n > 1 \end{cases} \tag{6}$$

After illumination enhancement, for all $y_f^1, y_f^2, ..., y_f^S$, we use a 3×3 convolution layer to generate feature $z_f^i = Conv_i(y_f^i)$. Finally, the different features $z_f^1, z_f^2, ..., z_f^S$ are concatenated to form the output of IMAConv.

$$z_f = \text{Concat}(z_f^1, z_f^2, ..., z_f^S), \tag{7}$$

Deformable Bi-directional Cross-Attention

As previously described and analyzed, the high-quality quantized features obtained from the codebook are not flawless. The structural warping and textural distortion leads to a more severe misalignment between high-quality VQ codebook features and original degraded features. Therefore, we propose the Deformable Bi-directional Cross-Attention (DBCA) to fuse high-quality VQ codebook features and degraded features.

Unlike the conventional cross-attention method (2021), our DBCA aims to integrate two different features using a bi-directional cross-attention mechanism and employs deformable convolutions to effectively correct the blurring degradation in the degraded feature. As shown in Figure 5, given the input decoder D and G features F_D and F_G, they are first mapped to corresponding $\mathbf{Q}_D = W_D^p \text{LN}(F_D)$ and $\mathbf{Q}_G = W_G^p \text{LN}(F_G)$ via normalization and linear layers. We further utilize linear layers to map these features to corresponding values $\mathbf{V}_D$ and $\mathbf{V}_G$. We reshape the aforementioned $\mathbf{Q}_D$, $\mathbf{Q}_G$, $\mathbf{V}_D$, and $\mathbf{V}_G$ into the shape of $(B, C, H * W)$ and fuse the two features using the following bi-directional cross-attention formula:

$$\text{A}_D = \text{Softmax}(\mathbf{Q}_D \mathbf{Q}_G^T / \sqrt{C}) \mathbf{V}_D, \tag{8}$$

$$\text{A}_G = \text{Softmax}(\mathbf{Q}_D \mathbf{Q}_G^T / \sqrt{C}) \mathbf{V}_G, \tag{9}$$

$$F_D^o = \gamma_D \text{A}_G + F_D, \tag{10}$$

$$F_G^o = \gamma_G \text{A}_D + F_G, \tag{11}$$

where $\text{Softmax}(\cdot)$ denotes the softmax function. A_D and A_G respectively represent the attention maps for feature D and feature G. γ_D and γ_G are trainable channel-wise scales and initialized with zeros for stabilizing training.

To better fuse the high-quality codebook prior feature into the degraded feature, we first generate an offset by concatenating the two output features. Then, we use the generated offset in the deformable convolution to distort the texture feature and effectively remove the blurry degradation, which can be formalized as follows:

$$\text{offset} = \text{LKConv}(\text{Concat}(F_D^o, F_G^o)), \tag{12}$$

$$F_{out} = \text{DeformConv}(F_D^o, \text{offset}), \tag{13}$$

where $\text{LKConv}(\cdot)$ and $\text{DeformConv}(\cdot)$ denote the 7×7 convolution and the deformable convolution, respectively.

Training Objectives of VQCNIR

The training objective of VQCNIR comprises four components: (1) pixel reconstruction loss $\mathcal{L}_{pix}$ that minimizes the distance between the outputs and the ground truth; (2) code alignment loss $\mathcal{L}_{ca}$ enforces the codes of the night images to be aligned with the corresponding ground truth; (3) perceptual loss $\mathcal{L}_{per}$ which operates in the feature space, aims to enhance the perceptual quality of the restored images; and (4) adversarial loss $\mathcal{L}_{adv}$ for restoring realistic textures.

Specifically, we adopt the commonly-used L1 loss in the pixel domain as the reconstruction loss, represented by:

$$\mathcal{L}_{pix} = ||x_h - VQCNIR(x_n)||_1, \tag{14}$$

where the x_h and x_n denote high-quality ground truth and night image, respectively. To improve the matching performance of codes for night images with codes for high-quality images. We adopt the L_2 loss to measure the distance, which can be formulated as:

$$\mathcal{L}_{ca} = ||z^e - z_{\mathbf{q}}^e||_2^2, \tag{15}$$

where z^e and $z_{\mathbf{q}}^e$ are the night image code and the ground truth code, respectively. The total training objective is the combination of the above losses:

$$\mathcal{L}_{VQCNIR} = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{ca}\mathcal{L}_{ca} + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv}, \tag{16}$$

where λ_{pix}, λ_{ca}, λ_{per}, and λ_{adv} denote the scale factors of each loss function, respectively.

Method	PSNR↑	SSIM↑	LPIPS↓
Zero-DCE → MIMO	17.68	0.542	0.510
LLFlow → Restormer	21.50	0.746	0.357
LLFlow → Uformer	21.51	0.750	0.350
MIMO → Zero-DCE	17.52	0.570	0.498
Restormer → LLFlow	21.89	0.772	0.347
Uformer → LLFlow	21.63	0.758	0.342
KinD++*	21.26	0.753	0.359
DeblurGAN-v2*	22.30	0.745	0.356
DMPHN*	22.20	0.817	0.301
MIMO*	22.41	0.835	0.262
Restormer*	23.63	0.841	0.247
LLFlow*	24.48	0.846	0.235
LEDNet*	<u>25.74</u>	<u>0.850</u>	<u>0.224</u>
Ours	**27.79**	**0.875**	**0.096**

Table 1: Quantitative evaluation on the LOL-Blur dataset. The symbol * indicates the network is retrained on the LOL-Blur dataset. The best and second-best values are indicated with Bold text and Underlined text respectively.

Method	MUSIQ↑	NRQM ↑	NIQE ↓
RUAS →MIMO	34.39	3.322	6.812
LLFlow → Restormer	34.45	5.341	4.803
LLFlow → Uformer	34.32	5.403	4.941
MIMO → Zero-DCE	28.36	3.697	6.892
Restormer → LLFlow	35.42	5.011	4.982
Uformer → LLFlow	34.89	4.933	5.238
KinD++*	31.74	3.854	7.299
DMPHN*	35.08	4.470	5.910
MIMO*	35.37	5.140	5.910
Restormer*	36.65	5.497	5.093
LLFlow*	34.87	5.312	5.202
LEDNet	<u>39.11</u>	<u>5.643</u>	<u>4.764</u>
Ours	**51.04**	**7.064**	**4.599**

Table 2: Quantitative evaluation on the Real-LOL-Blur dataset. The symbol * indicates the network is retrained on the LOL-Blur dataset. The best and second-best values are indicated with Bold text and Underlined text respectively.

Experiments

Dataset and Training Details

We train our VQCNIR network on the LOL-Blur dataset (2022), which consists of 170 sequences (10,200 pairs) of training data and 30 sequences (1,800 pairs) of test data. We use random rotations of 90, 180, 270, random flips, and random cropping to 256×256 size for the augmented training data. We train our network using Adam (2014) optimizer with β_1=0.9, β_2=0.99 for a total of 500k iterations. The mini-batch size is set to 8. The initial learning rate is set to 1×10^{-4} and adopts the MultiStepLR to adjust the learning rate progressively. We empirically set λ_{pix}, λ_{ca}, λ_{per}, and λ_{adv} to $\{1, 1, 1, 0.1\}$. All experiments are performed on a PC equipped with Intel Core i7-13700K CPU, 32G RAM, and the Nvidia RTX 3090 GPU with CUDA 11.2.

Results on LOL-Blur Dataset

In this section, we compare our proposed VQCNIR quantitatively and qualitatively with all the above methods on the LOL-Blur test set (2022). We use the two most widely evaluated metrics: PSNR and SSIM for a fair evaluation of all methods. In addition, we employ the LPIPS metric to evaluate the perceptual quality of the restored images.

Quantitative Evaluations. Table 1 shows the quantitative results of our method and other methods on the LOL-Blur dataset. As shown, our method outperforms state-of-the-art LEDNet by 2.05 dB and 0.025 in PSNR and SSIM, respectively. Also, our method demonstrates clear advantages over existing methods when evaluated by perceptual quality metrics. The effectiveness of our method is well evidenced.

Qualitative Evaluations. Figure 6 shows the visual effect of all the compared methods. As the figure shows, most methods are ineffective in removing the blurring effect in severely blurred regions, inevitably introducing artifacts into the restored image. In contrast, our method can effectively recover the correct texture features by using high-quality prior information. Therefore, these results provide sufficient evidence that the codebook prior proposed by our method is particularly suitable for the task of night image restoration.

Results on Real Dataset

To better illustrate the effectiveness of our method in the real scene, we compare our proposed VQCNIR with the above method quantitatively and qualitatively under the real Real-LOL-Blur dataset (2022). Since the real scene lacks a corresponding reference image to evaluate, three non-reference evaluation metrics were used for the evaluation: MUSIQ (2021), NRQM (2017), and NIQE (2012). The MUSIQ metric assesses mainly color contrast and sharpness, which is more appropriate for this task.

Quantitative Evaluations. Table 2 exhibits the quantitative results of our method and other methods on the Real-LOL-Blur test set. As shown in Table 2, our method achieves the highest NIQE and NRQM scores, indicating that the restored results of our method have better image quality and are consistent with human perception. Moreover, we have the highest MUSIQ, which means that our results are the best in terms of color contrast and sharpness.

Qualitative Evaluations. Figure 7 displays the visual comparison results for all evaluated methods. As evident from the figure, simple cascade deblurring and low-light enhancement techniques can cause issues such as overexposure and blurring of saturated areas in the image. Even the end-to-end method of retraining on the LOL-Blur dataset suffers from undesired severe artifacts and blurring. In contrast, our proposed VQCNIR outperforms these methods in terms of visual quality, demonstrating fewer artifacts and blurring. This improvement can be attributed to the successful integration of a high-quality codebook prior into the network, which assists in generating high-quality textures. The comparison results of a real-world image further demonstrate the superiority of our proposed method.

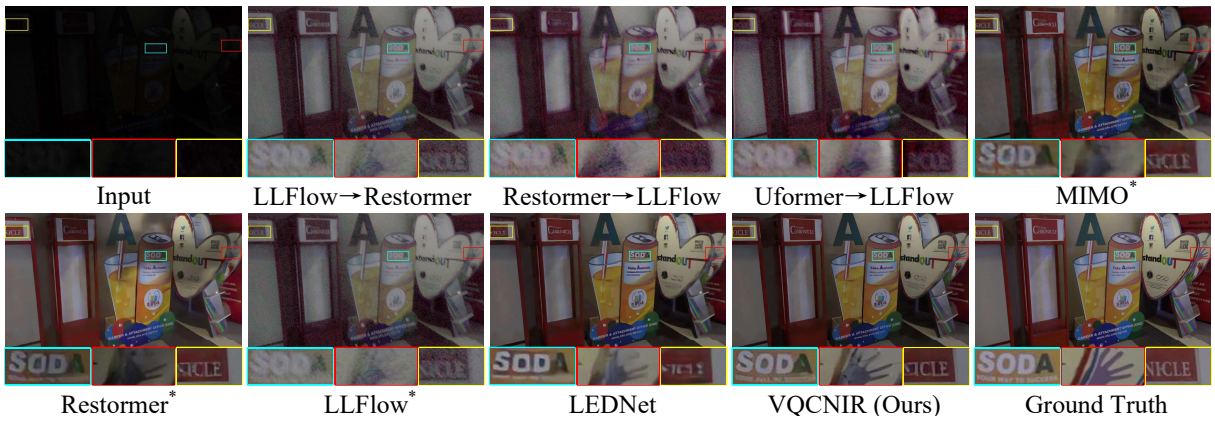


Figure 6: Visual comparison results on the LOL-Blur dataset (2022). The symbol * indicates the network is retrained on the LOL-Blur dataset. The proposed method produces visually more pleasing results. (Zoom in for the best view)

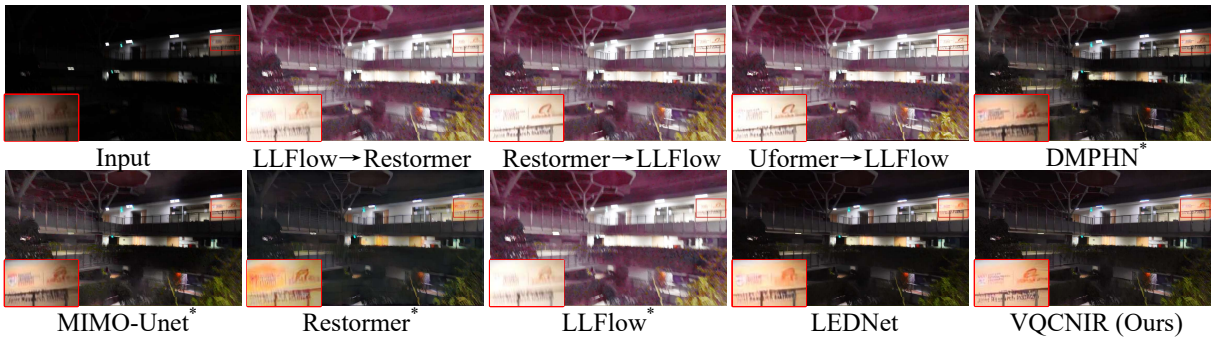


Figure 7: Visual comparison on the Real-LOL-Blur dataset (2022). The symbol * indicates the network is retrained on the LOL-Blur dataset. The proposed method produces visually more pleasing results. (Zoom in for the best view)

Models	Configuration			LOL-Blur	
	Decoder D	AIEM	DBCA	PSNR	SSIM
VQGAN				10.79	0.3028
Setting 1	✔			26.58	0.8486
Setting 2	✔		✔	26.89	0.8599
Setting 3	✔	✔		27.48	0.8692
VQCNIR	✔	✔	✔	27.79	0.8750

Table 3: Ablation studies of different components. We report the PSNR and SSIM values on the LOL-Blur dataset.

Ablation Study

In this section, we have implemented a series of ablation experiments to better validate the effectiveness of each of our proposed modules. To verify the effectiveness of our proposed operations, a series of ablation experiments are presented and the results are shown in Table 3. Initially, we use the VQGAN as our baseline model. Table 3 shows that VQGAN does not effectively address low light and blur degradation, since VQGAN is a codebook prior learned from high-quality natural images and is unable to correctly match degraded features. By designing corresponding parallel decoders, the network can then effectively use high-quality priors to assist in the reconstruction of degraded features. However, the illumination inconsistency between the degraded features and the codebook prior can prevent accurate matching of the high-quality prior features, leading to the occurrence of artifacts. Furthermore, the degraded features are at some distance from high-quality features. Therefore, AIEM and DBCA can be used to effectively improve network performance and image quality.

Conclusion

In this work, we introduce high-quality codebook priors and propose a new paradigm for night image restoration called VQCNIR. Through analysis, we discover that directly applying codebook priors can result in improper matching between degraded features and high-quality codebook features. To address this, we propose an Adaptive Illumination Enhancement Module (AIEM) and a Deformable Bi-directional Cross-Attention (DBCA) module, leveraging estimated illumination curves and bi-directional cross-attention. By fusing codebook priors and degraded features, VQCNIR effectively restores normal illumination and texture details from night images. Extensive experiments demonstrate the state-of-the-art performance of our method.

Acknowledgments

This work was supported by the Science and Technology Project of Guangzhou under Grant 202103010003, Science and Technology Project in key areas of Foshan under Grant 2020001006285.

References

Blau, Y.; Mechrez, R.; Timofte, R.; Michaeli, T.; and Zelnik-Manor, L. 2018. The 2018 PIRM challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Chen, C.; Shi, X.; Qin, Y.; Li, X.; Han, X.; Yang, T.; and Guo, S. 2022a. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1329–1338.

Chen, C.-F. R.; Fan, Q.; and Panda, R. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 357–366.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022b. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, 17–33. Springer.

Chen, L.; Zhang, J.; Lin, S.; Fang, F.; and Ren, J. S. 2021. Blind deblurring for saturated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6308–6316.

Chen, S.; Ye, T.; Bai, J.; Chen, E.; Shi, J.; and Zhu, L. 2023a. Sparse Sampling Transformer with Uncertainty-Driven Ranking for Unified Removal of Raindrops and Rain Streaks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13106–13117.

Chen, S.; Ye, T.; Liu, Y.; Bai, J.; Chen, H.; Lin, Y.; Shi, J.; and Chen, E. 2023b. CPLFormer: Cross-scale Prototype Learning Transformer for Image Snow Removal. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4228–4239.

Chen, S.; Ye, T.; Liu, Y.; Liao, T.; Jiang, J.; Chen, E.; and Chen, P. 2023c. MSP-former: Multi-scale projection transformer for single image desnowing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Chen, S.; Ye, T.; Xue, C.; Chen, H.; Liu, Y.; Chen, E.; and Zhu, L. 2023d. Uncertainty-Driven Dynamic Degradation Perceiving and Background Modeling for Efficient Single Image Desnowing. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4269–4280.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4641–4650.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Gu, Y.; Wang, X.; Xie, L.; Dong, C.; Li, G.; Shan, Y.; and Cheng, M.-M. 2022. 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*, 126–143. Springer.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1780–1789.

Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S.-M. 2022. Visual attention network. *arXiv preprint arXiv:2202.09741*.

Hu, Z.; Cho, S.; Wang, J.; and Yang, M.-H. 2014. Deblurring low-light images with light streaks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3382–3389.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30: 2340–2349.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer ence*.

Kupyn, O.; Martyniuk, T.; Wu, J.; and Wang, Z. 2019. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8878–8887.

Levin, A.; Weiss, Y.; Durand, F.; and Freeman, W. T. 2009. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE conference on computer vision and pattern recognition*, 1964–1971. IEEE.

Li, C.; Guo, C.; Han, L.; Jiang, J.; Cheng, M.-M.; Gu, J.; and Loy, C. C. 2022. Low-Light Image and Video Enhancement Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9396–9416.

Li, C.; Guo, C.; and Loy, C. C. 2021. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4225–4238.

Liu, Y.; Yan, Z.; Chen, S.; Ye, T.; Ren, W.; and Chen, E. 2023. Nighthazeformer: Single nighttime haze removal using prior query transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4119–4128.

Liu, Y.; Yan, Z.; Wu, A.; Ye, T.; and Li, Y. 2022. Nighttime image dehazing based on variational decomposition model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 640–649.

Lore, K. G.; Akintayo, A.; and Sarkar, S. 2017. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61: 650–662.

Ma, C.; Yang, C.-Y.; Yang, X.; and Yang, M.-H. 2017. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158: 1–16.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.

Shen, Z.; Wang, W.; Lu, X.; Shen, J.; Ling, H.; Xu, T.; and Shao, L. 2019. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5572–5581.

Tao, X.; Gao, H.; Shen, X.; Wang, J.; and Jia, J. 2018. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8174–8182.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6849–6857.

Wang, Y.; Wan, R.; Yang, W.; Li, H.; Chau, L.-P.; and Kot, A. 2022a. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2604–2612.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022b. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17683–17693.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*.

Ye, T.; Chen, S.; Liu, Y.; Chai, W.; Bai, J.; Zou, W.; Zhang, Y.; Jiang, M.; Chen, E.; and Xue, C. 2023. Sequential Affinity Learning for Video Restoration. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4147–4156.

Ye, T.; Zhang, Y.; Jiang, M.; Chen, L.; Liu, Y.; Chen, S.; and Chen, E. 2022. Perceiving and modeling density for image dehazing. In *European Conference on Computer Vision*, 130–145. Springer.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zhang, H.; Dai, Y.; Li, H.; and Koniusz, P. 2019. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5978–5986.

Zhang, K.; Luo, W.; Zhong, Y.; Ma, L.; Stenger, B.; Liu, W.; and Li, H. 2020. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2737–2746.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond brightening low-light images. *International Journal of Computer Vision*, 129: 1013–1037.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, 286–301.

Zhang, Y.; Zhang, J.; and Guo, X. 2019. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, 1632–1640.

Zheng, C.; Shi, D.; and Shi, W. 2021. Adaptive unfolding total variation network for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4439–4448.

Zhou, S.; Li, C.; and Change Loy, C. 2022. Lednet: Joint low-light enhancement and deblurring in the dark. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, 573–589. Springer.