The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Low-Light Image Enhancement via Generative Perceptual Priors

Han Zhou[1*], **Wei Dong**[1*], **Xiaohong Liu**[2†], **Yulun Zhang**[2], **Guangtao Zhai**[2], **Jun Chen**[1]

[1]McMaster University
[2]Shanghai Jiao Tong University
{zhouh115, dongw22, chenjun}@mcmaster.ca, {xiaohongliu, yulzhang, zhaiguangtao}@sjtu.edu.cn

Abstract

Although significant progress has been made in enhancing visibility, retrieving texture details, and mitigating noise in Low-Light (LL) images, the challenge persists in applying current Low-Light Image Enhancement (LLIE) methods to real-world scenarios, primarily due to the diverse illumination conditions encountered. Furthermore, the quest for generating enhancements that are visually realistic and attractive remains an underexplored realm. In response to these challenges, we introduce a novel **LLIE** framework with the guidance of **G**enerative **P**erceptual **P**riors (**GPP-LLIE**) derived from vision-language models (VLMs). Specifically, we first propose a pipeline that guides VLMs to assess multiple visual attributes of the LL image and quantify the assessment to output the global and local perceptual priors. Subsequently, to incorporate these generative perceptual priors to benefit LLIE, we introduce a transformer-based backbone in the diffusion process, and develop a new layer normalization (***GPP-LN***) and an attention mechanism (***LPP-Attn***) guided by global and local perceptual priors. Extensive experiments demonstrate that our model outperforms current SOTA methods on paired LL datasets and exhibits superior generalization on real-world data.

Code — https://github.com/LowLevelAI/GPP-LLIE

1 Introduction

Images captured in low-light conditions exhibit compromised quality, characterized by diminished visibility, reduced contrast, and loss of detail, which complicates both human observation and vision-based tasks (Huang et al. 2016; Dong, Zhou, and Xu 2018; Xu, Dong, and Zhou 2022). Enhancing these images is particularly challenging in real-world scenarios due to complex degradation caused by diverse illumination levels and elevated noise intensities.

Although substantial progress has been made for low-light image enhancement (LLIE) (Zhou et al. 2024; Li et al. 2023; Dong et al. 2024c), traditional methods, which rely on handcrafted descriptors(Fu et al. 2016; Jobsob, Zia-ur, and Woodell 1997; Shen and Gupta 2022) and existing deep learning approaches, which exploit the data fitting capability of neural networks (Yan et al. 2016; Yang et al. 2020; Lore, Adedotun, and Soumikm 2017), still suffer from poor adaptability in real-world scenarios. Image priors (*i.e.*, edge) (Xu, Wang, and Lu 2023; Deepanshu, Lal, and Parihar 2021; Kim et al. 2021; Zhu et al. 2020), semantic priors (Wu et al. 2023; Fan et al. 2020), and illumination maps (Wang et al. 2020) are commonly used for improving image quality. However, their effectiveness in generating realistic details is limited due to the difficulty of predicting robust priors from severely degraded inputs. Semantic priors are also constrained by the reliance on predefined semantic categories, which affects their ability to generalize in practical applications. Recent work (Luo et al. 2024) has attempted to fine-tune pre-trained VLMs to align the content embedding for degraded images to that for clean images. Yet, fine-tuning VLMs on limited data has a risk of over-fitting and demonstrates limited generalizability on unseen data.

On the other hand, several recent generative LLIE approaches (Jiang et al. 2023a; Zhou et al. 2023a; He et al. 2023; Yin et al. 2023) are proposed based on Diffusion Model (DM) (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021), which demonstrates impressive capacity to represent intricate distributions without succumbing to the mode collapse and training instability often associated with GANs. Despite the prevailing choice of convolutional U-Net network as the de facto backbone, (Peebles and Xie 2023) indicates that the inductive biases inherent in U-Net may not be pivotal for the remarkable performance of diffusion models and introduces the transformer-based diffusion model (DiT) with good scalability for image synthesis. However, directly employing DiT for LLIE is unfeasible: **(1)** The original DiT network is designed to output images with specific resolutions, whereas LLIE models typically process images of varying sizes; **(2)** The computational complexity of the Vision Transformer in DiT scales quadratically with the input size, which limits its applicability to high-resolution images; **(3)** Lacking prior information, the original DiT often demonstrates limited generalization to real-world images.

In view of the aforementioned challenges, we propose a novel **LLIE** framework with the guidance of **G**enerative **P**erceptual **P**riors (**GPP-LLIE**) derived from vision-language models (VLMs). Firstly, we propose incorporating external perceptual priors to help LLIE model to

[*]These authors contributed equally.
[†]Corresponding Author

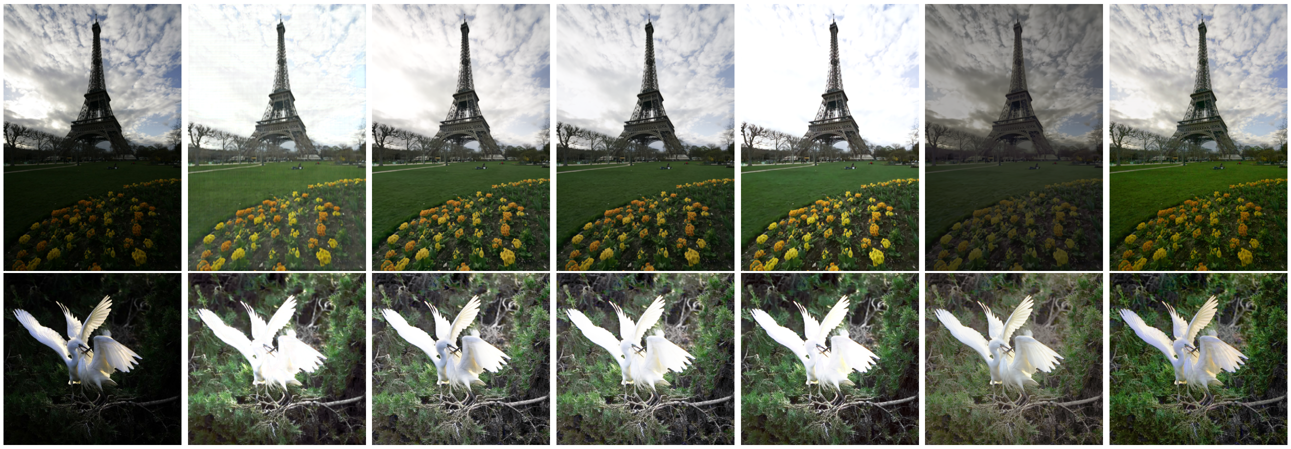

Figure 1: Visual comparisons on real-world datasets without Ground Truth. The images analyzed are sourced from MEF (Ma, Zeng, and Wang 2015) and NPE (Wang et al. 2013), respectively. Our method stands out by providing a **balanced enhancement effect**. Unlike other approaches, our method effectively enhances luminance, not only to reveal finer details but also preserve natural color tones. Notably, the ***cloud details*** in the sky, the ***structural integrity*** of the Eiffel Tower and the ***texture of flowers and branches*** are rendered with clarity and ***without over-exposure artifacts*** or unnatural coloration, distinctly surpassing others.

comprehend the varying attributes in LL images, thereby guiding the enhancement process. Different from previous strategies that extract features from severely degraded images, we develop a novel pipeline as shown in Fig. 2 to extract generative perceptual priors from VLMs, which are pre-trained on extensive datasets and exhibit proficiency in perceiving visual low-level attributes on unseen images. Specifically, we develop instruction prompts to guide VLMs to assess multiple attributes (*i.e.*, contrast, visibility, and sharpness) of LL images globally and locally. Then, these assessments are quantified into the global and local perceptual priors via our proposed sigmoid-based quantification strategy. Secondly, we develop an efficient transformer-based backbone with the guidance of generative perceptual priors in the diffusion process. Specifically, the global perceptual prior is employed to modulate the layer normalization (***GPP-LN***) and the local perceptual prior is introduced to guide the attention mechanism (***LPP-Attn***). Experimental results attest to the effectiveness of our proposed network across diverse real-world scenarios, yielding visually appealing and detail-rich enhancement results as shown in Fig. 1 In addition, we also apply our generative perceptual priors to other LLIE models and enhanced results further demonstrate the effectiveness of our proposed pipeline of extracting generative perceptual priors from VLMs. We highlight our contributions as follows:

(1) We introduce an **innovative pipeline** to acquire **generative perceptual priors** for LL images globally and locally based on pre-trained VLMs.

(2) With the guidance of global and local generative perceptual priors, we develop an efficient transformer based diffusion framework for LLIE (**GPP-LLIE**).

(3) We introduce global perceptual priors to modulate the layer normalization (**GPP-LN**) and leverage local perceptual priors to guide the attention mechanism (**LPP-Attn**) in our transformer block to benefit the enhancement process.

(4) Our method demonstrates SOTA performance on **various benchmark datasets** and exhibits **good generalization on real-world data**. Furthermore, our generative perceptual priors are applicable to help other LLIE models achieve enhanced outcomes.

2 Related Works

Diffusion-Based LLIE Compared to common deep learning architectures for image restoration (Liu et al. 2019, 2022; Dong et al. 2024b; Zhou et al. 2023b; Dong et al. 2024a; Ancuti et al. 2023, 2024; Vasluianu et al. 2024; Yin, Liu, and Liu 2019; Fu et al. 2024), due to its remarkable generative capabilities, diffusion model (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021) has emerged as a dominant force in recently proposed image-level generative models, finding widespread applications in image synthesis (Peebles and Xie 2023), inpainting (Rombach et al. 2022), and super-resolution (Sun et al. 2024). Diffusion model is firstly introduced for LLIE in Diff-Retinex (Yi et al. 2023) which employs two diffusion networks to reconstruct normal-light illumination and reflectance maps to produce the final result. Then, Reti-Diff (He et al. 2023) proposes to utilize the diffusion process in the latent space to alleviate computational costs and leverage a transformer network for detail refinement. CLEDiff (Yin et al. 2023) introduces a diffusion network conditioned on brightness level. PyDiff (Zhou et al. 2023a) adopts a strategy that progressively increases the resolution in the reverse diffusion process. However, all these methods utilize convolutional U-Nets architecture as the backbone for diffusion model. The untapped potential of Transformers, renowned for their scalability and effectiveness, remains unexplored in diffusion-based LLIE methods.

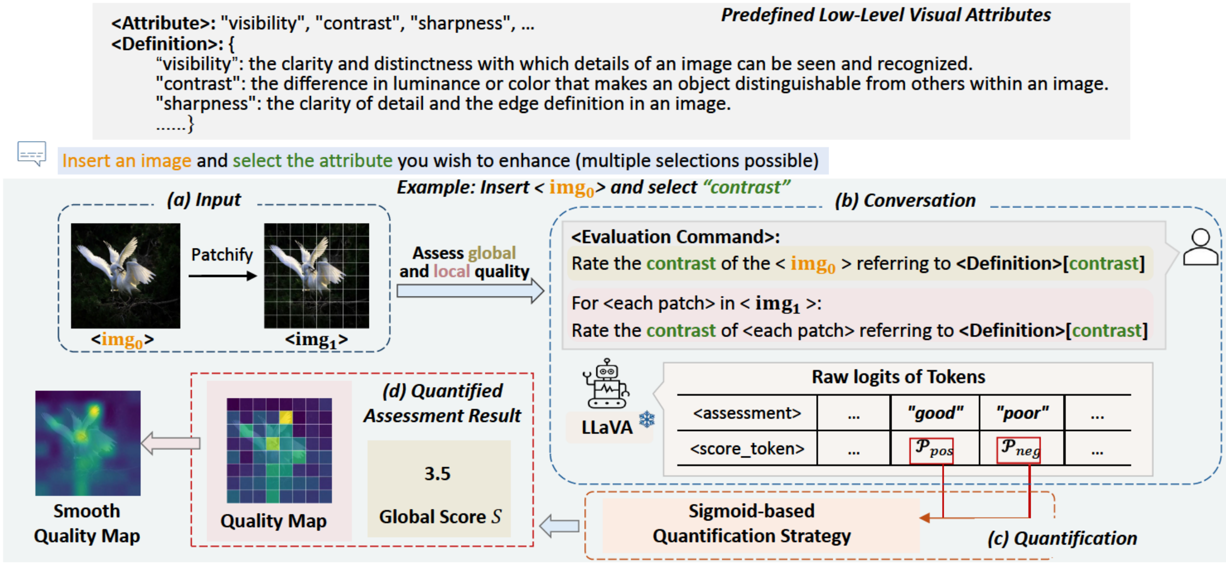

Figure 2: Our proposed pipeline for fine-grained **generative perceptual priors extraction** from pre-trained VLMs. (a) The input image is patchified into several non-overlapped patches. (b) Conversation with VLMs. We develop evaluation commands to guide VLMs to assess the image globally and locally regarding the selected attribute. (c) Based on the output of VLMs, we design the sigmoid-based quantification strategy. (d) Our extracted global and local generative priors.

Image Restoration with Vision-Language Guidance Pre-trained with large-scale image-text datasets, vision-language models (*i.e.*, CLIP (Radford et al. 2021)) demonstrate impressive zero-shot capability and have been applied to image restoration in recent years. For example, CLIP-LIT (Liang et al. 2023) introduces learnable prompts for unsupervised backlit image enhancement. DACLIP (Luo et al. 2024) designs an additional image controller to separate degradation and image features from the frozen CLIP latent space for universal image restoration. These methods mainly concentrate on applying VLMs to generate captions of image content. However, accurately describing the content of low-light, complex scene images presents a challenge. Instead, based on pre-trained VLMs, we propose an innovative pipeline to extract generative perceptual priors regarding the low-level attributes of LL images. Moreover, we incorporate these priors into our proposed transformer-based diffusion network to enhance the performance in LLIE.

3 Method

The main focus of this work is to extract generative perceptual priors that well represent visual attributes of LL images and to develop LLIE models guided by these priors to generate realistic and visually attractive enhancement results. The overall framework is illustrated in Fig. 3.

In this section, we first discuss the motivation of employing guidance derived from vision language models (VLMs) for LLIE task (Sec. 3.1). Then, we propose an innovative pipeline that guides VLMs to assess the visual attributes of LL images globally and locally and then extract the perceptual priors by introducing the sigmoid-based quantification strategy (Sec. 3.2). Moreover, we develop a transformer based diffusion framework and we incorporate these priors to guide the reverse diffusion process (Sec. 3.3).

3.1 Motivation of Utilizing VLMs Guidance

Although recent methods in low-light image enhancement (LLIE) have shown improved performance, they often yield unbalanced results with over-exposure artifacts when applied to real-world images, which frequently differ in lighting conditions from the training datasets (see Fig. 1). These results underscore a general inability of current LLIE methods to adaptively enhance images under varied illumination conditions. Thus, enabling models to autonomously ***perceive*** and adapt to various visual distortions is of vital importance. Inspired by the recently demonstrated capabilities of emergent Vision-Language Models (VLMs) in low-level visual perception and understanding (Wu et al. 2024), we aim to explore the potential of utilizing these perceptual abilities of VLMs to facilitate LLIE tasks.

3.2 Generative Perceptual Priors from VLMs

VLMs are usually trained with millions of text-image pairs and demonstrate remarkable zero-shot capabilities in generating aligned understandings between texts and images. Therefore, it is essentially promising to utilize the prior information inherent in VLMs to help LLIE models make more appropriate decisions during the restoration process. However, the VLMs employed in recent image restoration works (Luo et al. 2024; Sun et al. 2024) are primarily focused on understanding the semantic content of images, yet they lack precise representation of visual details. Moreover, accurately describing the content of complex LL images is quite challenging. In contrast, the VLMs employed in our work is LLaVA (Liu et al. 2023a), which is further fine-tuned

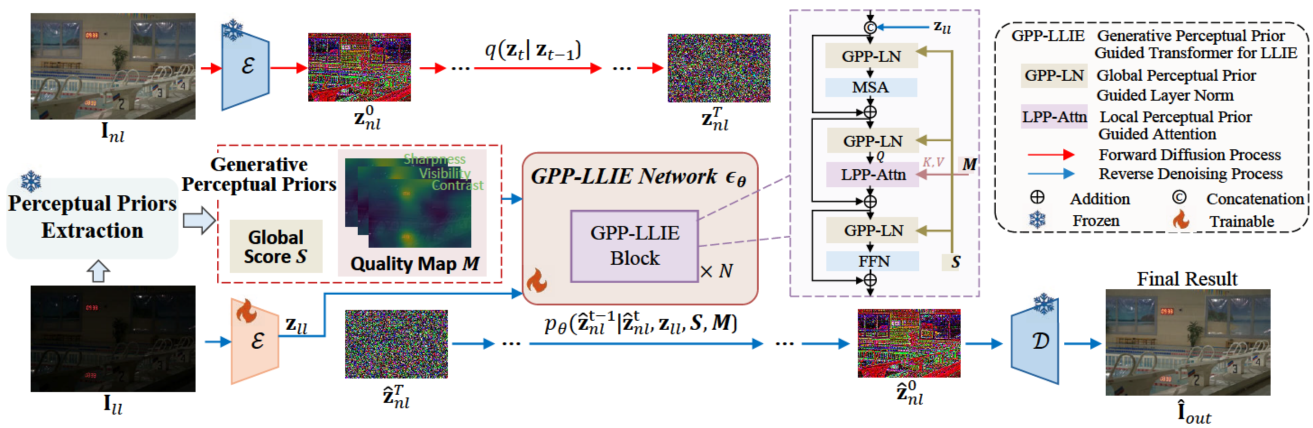

Figure 3: The overall framework of our proposed **GPP-LLIE** method. We first employ the encoder $\mathcal{E}$ to convert the NL image $\mathbf{I}_{nl}$ and LL image $\mathbf{I}_{ll}$ into latent space denoted as $\mathbf{z}_{nl}^0$ and $\mathbf{z}_{ll}$. Then, the forward diffusion process is applied upon $\mathbf{z}_{nl}^0$. In order to leverage the prior information to guide the reverse diffusion process, the $\mathbf{I}_{ll}$ is sent to our proposed pipeline for perceptual priors extraction. With the guidance of global perceptual score S, local quality map $\boldsymbol{M}$, and LL feature $\mathbf{z}_{ll}$, we develop a transformer-based network ϵ_θ and gradually transform the randomly sampled Gaussian noise $\hat{\mathbf{z}}_{nl}^T$ into a clear NL latent feature $\hat{\mathbf{z}}_{nl}^0$. Finally, the restored feature $\hat{\mathbf{z}}_{nl}^0$ is fed into the decoder $\mathcal{D}$ to generate the final enhancement $\hat{\mathbf{I}}_{out}$.

with 200K instruction-response pairs related to low-level visual aspects in Q-instruct (Wu et al. 2024). In this paper, we introduce a new pipeline to employ LLaVA in LLIE: we design text prompts to guide LLaVA to assess multiple visual attributes of LL images. In addition, different from the text/image embedding in previous methods (Sun et al. 2024; Luo et al. 2024), we introduce the quantification strategy to output the quantified global assessment and local quality map as perceptual priors for LLIE. Our pipeline for perceptual priors extraction is shown as Fig. 2.

Fine-grained Low-Level Visual Assessment. Different from simply evaluating the overall quality of image, we offer several low-level visual attributes **<Attribute>** for selection and provide corresponding definitions **<Definition>** to help VLMs better understand the assessment task. Specifically, given an image **<img$_0$>**, we can select the attribute from **<Attribute>** for assessment. For example, as LLIE is to enhance the contrast, visibility, and sharpness of LL images, we can sequentially evaluate these attributes in LL inputs.

Besides the global assessment, considering the varying contrast, visibility, and sharpness within LL images, we also propose to extract local assessment for fine-grained enhancement. Specifically, the input **<img$_0$>** is patchified into several non-overlapped patches **<img$_1$>** and each patch is fed into LLaVA to acquire the local assessment. Therefore, the overall evaluation command during the conversation is defined as the **<Evaluation Command>** in Fig. 2.

Sigmoid-based Quantification Strategy. With the input image and evaluation command, we observe several issues in the probabilities of tokens generated by LLaVA: (1) Tokens (*i.e.*, ***"The"***) that have the highest probability are meaningless. (2) Compared to a single token, the difference between two opposite logits (*i.e.*, ***"good"*** and ***"poor"***) is more aligned with human perception. These observations motivate us to explore additional strategies to generate quantified output well-aligned with human opinions for LLIE.

To start with, we design our strategy based on tokens with contextual information rather than those with the highest probability. Moreover, our strategy is designed based on the probability of ***"good"*** and ***"poor"***, which can be considered as positive and negative assessments of the attributes of LL images. This strategy aligns more closely with human perceptual system, as assessments of images typically include both positive and negative evaluations. Therefore, the quantified global score S is calculated as $S = (1+e^{-(\mathcal{P}_{pos}-\mathcal{P}_{neg})/\alpha})^{-1}$, where a sigmoid operation modulates the difference between the probabilities of ***"good"*** and ***"poor"*** (denoted as $\mathcal{P}_{pos}$ and $\mathcal{P}_{neg}$), α is the modulation scalar and is set as 3 in this work. Similarly, we calculate the score for each patch in **<img$_1$>** and then obtain the quality map $\boldsymbol{M}$. Specific to LLIE task, three attributes ("contrast", "visibility", and "sharpness") are evaluated and the average global score and concatenated quality map are introduced as perceptual prior guidance in our proposed LLIE model.

3.3 Generative Perceptual Prior Guided Diffusion Transformer

To achieve enhanced generalizability on unseen real-world images, we build our LLIE model based on the Diffusion Transformer (DiT) network (Peebles and Xie 2023), which shares similar architecture with Vision Transformers (ViT) and presents good scalability properties. However, the DiT is originally designed for image synthesis at specific resolutions (*i.e.*, 256×256 or 512×512) and the computational complexity of ViT is quadratic to the input size. Evidently, the original DiT is infeasible for LLIE task, as LLIE models usually process LL images with variable sizes and sometimes large resolutions. To this end, we introduce a transformer-based backbone in the diffusion process, which is suitable for LLIE and contains special designs for incorporating external generative perceptual priors.

Methods	LOL				LOL-v2-real				LOL-v2-syn			
	FID ↓	LPIPS↓	DISTS↓	PSNR↑	FID ↓	LPIPS↓	DISTS↓	PSNR↑	FID ↓	LPIPS↓	DISTS↓	PSNR↑
KinD (Zhang et al. 2019)	78.28	0.157	0.110	19.03	95.02	0.151	0.112	18.05	97.32	0.263	0.180	16.81
MIRNet (Zamir et al. 2020)	71.16	0.131	0.105	24.14	82.25	0.138	0.116	20.02	40.18	0.102	0.126	21.94
DRBN (Yang et al. 2020)	85.57	0.155	0.108	19.86	94.22	0.147	0.119	20.29	28.74	0.085	0.097	23.22
SNR-Net (Xu et al. 2022)	66.47	0.115	0.092	24.70	68.56	0.120	0.095	21.48	19.96	0.056	0.063	24.14
URetinex-Net (Wu et al. 2022)	85.59	0.121	0.096	21.33	76.74	0.144	0.107	21.16	33.25	0.075	0.087	24.73
LLFlow (Wang et al. 2022)	65.17	0.113	0.094	25.19	70.68	0.135	0.102	26.53	20.24	0.044	0.056	26.23
RQLLIE (Liu et al. 2023b)	53.32	0.121	0.086	25.24	68.89	0.142	0.102	22.37	16.96	0.044	0.053	25.94
Retinexformer (Cai et al. 2023)	72.38	0.131	0.106	25.16	79.58	0.171	0.115	22.80	22.78	0.059	0.066	25.67
CUE (Zheng et al. 2023)	69.83	0.224	0.141	21.86	67.05	0.133	0.112	21.19	31.33	0.076	0.083	24.41
CLEDiff (Yin et al. 2023)	86.94	0.164	0.108	25.51	82.27	0.183	0.118	22.68	18.58	0.064	0.080	27.38
LL-SKF (Wu et al. 2023)	59.47	0.105	0.084	26.80	57.84	0.111	0.084	28.45	21.58	0.040	0.063	29.11
Reti-Diff (He et al. 2023)	49.14	0.105	0.082	25.35	43.18	0.087	0.069	22.97	13.26	0.038	0.051	27.53
PyDiff (Zhou et al. 2023a)	48.28	0.099	0.079	27.09	44.27	0.094	0.072	28.77	15.69	0.040	0.055	29.27
DiffLL (Jiang et al. 2023b)	48.11	0.118	0.091	26.33	45.36	0.089	0.064	28.66	13.66	0.038	0.047	28.87
Ours	**36.73**	**0.081**	**0.063**	**27.51**	**26.78**	**0.055**	**0.047**	**29.23**	**9.74**	**0.031**	**0.039**	**30.17**

Table 1: Quantitative comparisons on LOL, LOL-v2-real, and LOL-v2-synthetic, our method achieves superior performance compared to current SOTA methods. These numbers are obtained either by using their released weights or by re-training their models. [Key: **Best**, Second Best, ↑ (↓): Larger (smaller) values leads to better performance]

Overview. The overall framework of our generative perceptual priors guided diffusion transformer is shown as Fig. 3. Given a paired normal-light (NL) image $\mathbf{I}_{nl} \in \mathbb{R}^{H\times W\times 3}$ and low-light (LL) image $\mathbf{I}_{ll} \in \mathbb{R}^{H\times W\times 3}$, an encoder $\mathcal{E}$ is employed to extract their latent representations $\mathbf{z}_{nl}^0 \in \mathbb{R}^{\frac{H}{f}\times\frac{W}{f}\times d}$ and $\mathbf{z}_{ll} \in \mathbb{R}^{\frac{H}{f}\times\frac{W}{f}\times d}$, where H, W, d, and f represent the image height, image width, the hidden dimension, and the down-sampling factor of $\mathcal{E}$. Then, the forward diffusion process is applied upon $\mathbf{z}_{nl}^0$ and its noised version is denoted as $\mathbf{z}_{nl}^T$. For the reverse demonising process, we gradually transform the randomly sampled Gaussian noise $\hat{\mathbf{z}}_{nl}^T$ into a clear NL latent feature $\hat{\mathbf{z}}_{nl}^0$ step by step. For each step t, besides the latent representation for LL image $\mathbf{z}_{ll}$, we also incorporate the generative perceptual prior extracted from LLaVA (shown in Fig.2) as the guidance in our proposed GPP-LLIE network ϵ_θ. Finally, the restored feature $\hat{\mathbf{z}}_{nl}^0$ is fed into the decoder $\mathcal{D}$ to produce the final result $\hat{\mathbf{I}}_{out}$.

GPP-LLIE Network. Our GPP-LLIE network ϵ_θ is shown as Fig. 3, highlighting its several unique characteristics:

Concat-and-Remove Strategy: Within each GPP-LLIE block, we first concatenate the LL feature $\mathbf{z}_{ll}$ to the input to introduce the LL information into the reverse diffusion process to enhance the fidelity. While at the end of the block, we remove the latter half of the channels, enabling the concatenation of the $\mathbf{z}_{ll}$ at the start of the next GPP-LLIE block.

Global Perceptual Prior Guided Layer Norm (GPP-LN): To effectively integrate the global score S derived from VLMs into our GPP-LLIE block, we modulate the layer normalization process. This modulation, driven by the scale and shift parameters (γ and β) influenced by S, optimizes the normalization process to better reflect the perceptual insights provided by the global perceptual prior. Given an input feature $\mathbf{z}_{in}$, the output of our GPP-LN operation is calculated by: $\mathbf{z}_{out} = \gamma \cdot \mathrm{LN}(\mathbf{z}_{in}) + \beta$, where $\gamma, \beta = \mathrm{MLP}(S)$.

Local Perceptual Prior Guided Attention (LPP-Attn): To reduce the huge computational cost caused by spatial self-attention mechanism, we calculate the attention map along the channel dimension in our GPP-LLIE block. Moreover, beside the MSA, we also develop another channel attention mechanism guided by the local quality map $\boldsymbol{M}$. Specifically, *query* element is calculated upon the input feature, while the calculation of *key* and *value* elements are guided by the local perceptual prior $\boldsymbol{M}$. Moreover, to facilitate the application of our LLIE model to LL images of diverse sizes, we remove the positional embedding from the Vision Transformer. Instead, the spatial positional embedding are learned with the guidance of the local perceptual prior $\boldsymbol{M}$.

4 Experiments

4.1 Experiment Settings

Training and Diffusion. Our model is trained using the AdamW optimizer with the total training iterations of 1.5M for all datasets and the learning rate is set to 10^{-4}. Each training input is cropped into 320×320, and the batch size is set to 16. We use horizontal flips and rotations for data augmentation. For the diffusion process, the total timesteps for training is set to $1,000$ during the training, and we use 25 steps to accelerate sampling process for the inference.

Low Light Datasets and Metrics. We conduct experiments on various low light datasets including LOL (Wei et al. 2018), LOL-v2-real, and LOL-v2-synthetic (Yang et al. 2021). Specifically, we train our model using 485, 689, and 900 LL-NL pairs on LOL, LOL-v2-real, and LOL-v2-synthetic datasets, and other 15, 100, and 100 images are used for evaluation. Moreover, we also test the generalization of our method on several real-world datasets without ground truth images including MEF (Ma, Zeng, and Wang 2015), LIME (Guo 2016), DICM (Lee, Lee, and Kim 2013), and NPE (Wang et al. 2013). Metrics for paired datasets includes Fréchet Inception Distance (FID) (Heusel et al. 2017), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), Deep Image Structure and Texture Similarity (DISTS) (Ding et al. 2020), and Peak Signal-to-Noise Ratio (PSNR). Besides, we adopt a no-reference metric Natural Image Quality Evaluator (NIQE) (Mittal et al. 2012) for unpaired real data evaluation.

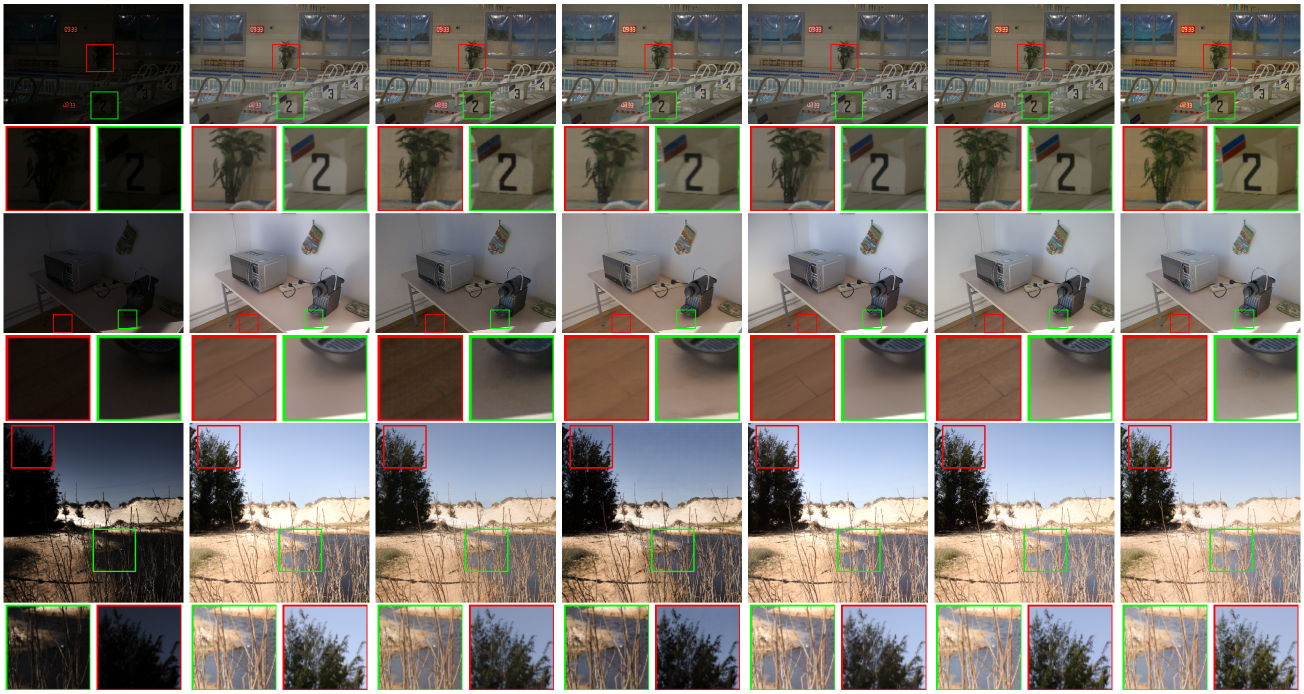

Figure 4: Visual comparisons on paired dataset. The images are sourced respectively from LOL (row 1), LOL-v2-real (row 2), and LOL-v2-syn (row3). Previous methods often result in overly smoothed images, consequently obscuring pivotal textural details. In contrast, our method yields sharper images while retaining the delicate details. For instance, our method maintains the structural complexities of the foliage and branches within the potted plant (first row). In second row, the grain on the wooden floor surface, as well as its edge contours, have been well-preserved. Similarly, in the natural landscapes, our method excels at enhancing the clarity of twig edges and maintaining color consistency.

4.2 Performance on LLIE

Quantitative Results on Paired Dataset. Tab. 1 summarizes the quantitative comparisons between our method with current SOTA methods. Our method achieves the superior performance on 3 commonly-used benchmarks, highlighting its advance and perfect generalization. Notably, the FID scores of our method surpass the best of SOTA methods by $\mathbf{23.6\%}$, $\mathbf{37.4\%}$, and $\mathbf{26.5\%}$ on the LOL, LOL-v2-real, and LOL-v2-synthetic, respectively. Moreover, our LPIPS values significantly outperforms the PyDiff by $\mathbf{18.4\%}$. These numbers demonstrate the superior perceptual quality of our enhancement and prove the effectiveness of generative perceptual priors in our method. Besides, our leading DISTS and PSNR scores show the satisfactory ability of our model in recovering texture details and maintaining fidelity.

Qualitative Results on Paired Datasets. We present the enhanced images of different methods in Fig. 4. Our appealing and realistic enhancement results demonstrate our method can generate images with pleasant illumination, correct color retrieval, and enhanced texture details. For example, the rich structural details of potted plants in row 1, the well preserved wooden floor surface and its edge contours in row 2, and the enhanced twig edge in row 3. In contrast, previous methods tend to output blurry results without rich textures (Retinexformer and PyDiff in row 2) and struggle to preserve color fidelity and illumination harmonization (LLFlow and RQ-LLIE in row 1).

Methods	MEF	LIME	DICM	NPE	Mean
SNR-Net (Xu et al. 2022)	4.14	5.51	4.62	4.36	4.60
LLFlow (Wang et al. 2022)	3.92	5.29	3.78	4.16	3.98
LL-SKF (Wu et al. 2023)	4.03	5.15	3.70	4.08	3.92
RQLLIE (Liu et al. 2023b)	4.21	4.86	4.02	**4.03**	4.13
CLEDiff (Yin et al. 2023)	5.27	5.00	4.42	4.57	4.57
PyDiff (Zhou et al. 2023a)	4.24	4.88	4.32	4.38	4.37
Ours	**3.55**	**4.24**	**3.58**	4.05	**3.67**

Table 2: Quantitative comparisons on real-world datasets in terms of NIQE. These numbers are obtained by testing with their released LOL weights. [Key: **Best**, Second Best]

Performances on Real-world Dataset. To further verify the generalization of our method, we extend our evaluations to real-world low light datasets. Notably, we select several methods that perform well in Tab. 1 and we use the weights of various methods trained on LOL training split to ensure fairness. The quantitative comparisons are presented as Tab. 2, where the NIQE metric is employed. It can be seen that our model achieve the best performance on the first 3 datasets and the second best result on NPE, thereby significantly outperforming other methods across real-world data.

Moreover, visual comparisons are reported in Fig. 1 and

Figure 5: Visual comparisons on real-world datasets (DICM (Lee, Lee, and Kim 2013)). Our method adeptly handles the ***diverse and uneven*** illumination levels present in the original image. It effectively enhances brightness and contrast while ***avoiding the overexposure*** of original bright areas and ***maintaining natural coloration***, generating visually appealing results.

Model	FID ↓	LPIPS↓	DISTS↓	PSNR↑
Ours	**36.73**	**0.081**	**0.063**	**27.51**
Variant 1	49.83	0.103	0.084	26.88
Variant 2	47.18	0.100	0.081	27.06

Table 3: By integrating the local perceptual prior using LPP-Attn mechanism, GPP-LLIE shows enhanced performance in terms of FID and LPIPS. Besides, our LPP-Attn performs better than spatial feature transform applied in StableSR.

Model	FID ↓	LPIPS↓	DISTS↓	PSNR↑
Variant 1	**49.83**	**0.103**	**0.084**	**26.88**
Variant 3	61.36	0.113	0.095	26.14
Variant 4	58.36	0.111	0.092	26.30

Table 4: Unlike GPP-LN, removing the global perceptual score or directly adding it to $\hat{\mathbf{z}}_{nl}^{T}$ yields quite poor results.

Fig. 5, where previous methods tend to output overexposure results or struggle to preserve details. In contrast, our enhanced images are more realistic and appealing and our method can flexibly enhance images with various illuminations. Specifically, for outdoor images with uneven illumination distribution, our method enhances the brightness while simultaneously preventing local overexposure, thereby preserving more details (*i.e.*, the enhanced architectural details with sharper edges and preserved details in row 1 of Fig. 5, and the improved contrast in the sky and cloud area in row 2 of Fig. 5 . These vivid and natural enhanced images, together with quantitative numbers reported in Tab. 1 and Tab. 2, demonstrate the superior effectiveness and generalization of our method compared to the current SOTA.

4.3 Ablation Study

To analyze the contribution of each component in our method, we conduct extensive ablation studies.

Local Perceptual Prior and LPP-Attn. To study the importance of local perceptual prior and our proposed LPP-Attn, we remove these two parts from our model and denote the remaining network as Variant 1. Tab. 3 reports the quantitative performance of Variant 1 on LOL dataset, which still presents competitive enhancement performance on all measured metrics compared to the current SOTA in Tab. 1. However, compared to Variant 1, our full GPP-LLIE shows significantly enhanced FID and LPIPS scores by integrating the local perceptual prior using our proposed LPP-Attn mechanism. Besides, we introduce the Variant 2 that replaces our LPP-Attn with spatial feature transform layer (Wang et al. 2018) applied in StableSR (Wang et al. 2023), our LPP-Attn mechanism achieves 22%, 19%, and 22% lower values in FID, LPIPS, and DISTS respectively. These comparisons manifest the importance of LPP-Attn and the local perceptual prior generated by our proposed extraction pipeline.

Global Perceptual Prior and GPP-LN. Based on Variant 1, we implement several adaptations to illustrate the effectiveness of our extracted global perceptual prior and the corresponding GPP-LN. (1) We remove global perceptual prior and GPP-LN and denote the remaining network as Variant 3. (2) We remove GPP-LN operation and directly add the global score to the noised latent feature $\hat{\mathbf{z}}_{nl}^{T}$. This framework is referenced as Variant 4. The quantitative results of Variant 3 and Variant 4 are reported in Tab. 4. The discernible performance disparity between Variant 3 and Variant 1 (*i.e.*, 23% higher FID and 10% higher in LPIPS) illustrates the superiority of the global perceptual score and our proposed GPP-LN. Besides, compared to Variant 4, the GPP-LN employed in Variant 1 better incorporates the prior information and guides our model achieve better performance (*i.e.*, 14.6% lower in FID and 7.2% lower in LPIPS).

5 Conclusion

To achieve adaptive and realistic enhancement in real-world scenarios, we introduce a novel LLIE framework (**GPP-LLIE**) guided by generative perceptual priors. We firstly develop a pipeline for generative perceptual priors extraction based on pre-trained VLMs. Specifically, we design several text prompts to guide VLMs to assess low-level attributes of low-light images globally and locally. We also introduce a sigmoid-based quantification strategy to output the global score and local quality map. Moreover, we introduce a transformer network as the backbone for the diffusion process, where we design the global perceptual priors modulated layer normalization and local perceptual priors guided attention mechanism to guide the enhancement. We evaluate our method on 3 paired and 4 real-world datasets, demonstrating superior performance and good generalization.

References

Ancuti, C. O.; Ancuti, C.; Vasluianu, F.-A.; Timofte, R.; et al. 2023. NTIRE 2023 HR NonHomogeneous Dehazing Challenge Report. In *CVPRW*.

Ancuti, C. O.; Ancuti, C.; Vasluianu, F.-A.; et al. 2024. NTIRE 2024 Dense and nonhomogeneous dehazing challenge report. In *CVPRW*.

Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; and Zhang, Y. 2023. Retinexformer: One-stage Retinex-based Transformer for Low-light Image Enhancement. In *ICCV*.

Deepanshu, R.; Lal, K. J.; and Parihar, A. S. 2021. Edge guided low-light image enhancement. In *Proceedings of the International Conference on Intelligent Computing and Control Systems*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image quality assessment: Unifying structure and texture similarity. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dong, W.; Zhou, H.; Tian, Y.; Sun, J.; Liu, X.; Zhai, G.; and Chen, J. 2024a. ShadowRefiner: Towards Mask-free Shadow Removal via Fast Fourier Transformer. In *CVPRW*.

Dong, W.; Zhou, H.; Wang, R.; Liu, X.; Zhai, G.; and Chen, J. 2024b. DehazeDCT: Towards Effective Non-Homogeneous Dehazing via Deformable Convolutional Transformer. In *CVPRW*.

Dong, W.; Zhou, H.; and Xu, D. 2018. A New Sclera Segmentation and Vessels Extraction Method for Sclera Recognition. In *2018 10th International Conference on Communication Software and Networks (ICCSN)*.

Dong, W.; Zhou, H.; Zhang, Y.; Liu, X.; and Chen, J. 2024c. ECMamba: Consolidating Selective State Space Model with Retinex Guidance for Efficient Multiple Exposure Correction. *arXiv preprint arXiv:2410.21535*.

Fan, M.; Wang, W.; Yang, W.; and Liu, J. 2020. Integrating Semantic Segmentation and Retinex Model for Low Light Image Enhancement. In *Proceedings of the ACM Conference on Multimedia*.

Fu, K.; Peng, Y.; Zhang, Z.; Xu, Q.; Liu, X.; Wang, J.; and Zhai, G. 2024. AttentionLut: Attention Fusion-based Canonical Polyadic LUT for Real-time Image Enhancement. *arXiv preprint arXiv:2401.01569*.

Fu, X.; Zeng, D.; Huang, X.; Xiao-Ping, Z.; and Xinghao, D. 2016. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*.

Guo, X. 2016. LIME: A method for low-light image enhancement. In *Proceedings of the ACM Conference on Multimedia*.

He, C.; Fang, C.; Zhang, Y.; Li, K.; Tang, L.; You, C.; Xiao, F.; Guo, Z.; and Li, X. 2023. Reti-Diff: Illumination Degradation Image Restoration with Retinex-based Latent Diffusion Model. *arXiv preprint arXiv:2311.11638*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in neural information processing systems*.

Huang, G.; Wang, X.; Wu, W.; Zhou, H.; and Wu, Y. 2016. Real-time lane-vehicle detection and tracking system. In *Chinese Control and Decision Conference (CCDC)*.

Jiang, H.; Luo, A.; Fan, H.; Han, S.; and Liu, S. 2023a. Low-light Image Enhancement with wavelet-based Diffusion Models. In *ACM Transactions on Graphics*.

Jiang, H.; Luo, A.; Han, S.; Fan, H.; and Liu, S. 2023b. Low-Light Image Enhancement with Wavelet-based Diffusion Models. In *Siggraph Asia*.

Jobsob, D.; Zia-ur, R.; and Woodell, G. 1997. A multiscale retinex for bridging the gap between color images and the human observation of scenes. In *IEEE Transactions on Image Processing*.

Kim, E.; Lee, S.; Park, J.; and Choi, S. 2021. Deep edge-aware interactive colorization against color-bleeding effects. In *ICCV*.

Lee, C.; Lee, C.; and Kim, C.-S. 2013. Contrast enhancement based on layered difference representation of 2D histograms. In *IEEE Transactions on Image Processing*.

Li, W.; Wu, G.; Wang, W.; Ren, P.; and Liu, X. 2023. FastLLVE: Real-Time Low-Light Video Enhancement with Intensity-Aware Lookup Table. In *Proceedings of the ACM Conference on Multimedia*.

Liang, Z.; Li, C.; Zhou, S.; Feng, R.; and Loy, C. C. 2023. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, H.; Li, C.; Wu, Q.; ; and Lee, Y. J. 2023a. Visual instruction tuning. In *NeurIPS*.

Liu, X.; Ma, Y.; Shi, Z.; and Chen, J. 2019. GridDehazeNet: Attention based Multi-Scale Network for Image Dehazing. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, X.; Shi, Z.; Wu, Z.; Chen, J.; and Zhai, G. 2022. GridDehazeNet+: An Enhanced Multi-Scale Network with Intra-Task Knowledge Transfer for Single Image Dehazing. In *IEEE Transactions on Intelligent Transportation Systems*.

Liu, Y.; Huang, T.; Dong, W.; Wu, F.; Li, X.; and Shi, G. 2023b. Low-Light Image Enhancement with Multi-stage Residue Quantization and Brightness-aware Attention. In *ICCV*.

Lore, K. G.; Adedotun, A.; and Soumikm, S. 2017. LLNet: A deep autoencoder approach to natural low-light image enhancement. In *Pattern Recognition*.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; and et.al. 2024. Controlling Vision-Language Models for Universal Image Restoration. In *ICLR*.

Ma, K.; Zeng, K.; and Wang, Z. 2015. Perceptual quality assessment for multi-exposure image fusion. In *IEEE Transactions on Image Processing*.

Mittal, A.; Soundararajan, R.; Bovik, A. C.; and et al. 2012. Making a 'completely blind' image quality analyzer. In *IEEE Signal Processing Letters*.

Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. In *ICCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Shen, Z.; and Gupta, G. 2022. Semantic-guided zero-shot learning for low-light image/video enhancement. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

Sun, H.; Li, W.; Liu, J.; Chen, H.; Pei, R.; Zou, X.; Yan, Y.; and Yang, Y. 2024. CoSeR: Bridging Image and Language for Cognitive Super-Resolution. In *CVPR*.

Vasluianu, F.-A.; Seizinger, T.; Zhou, Z.; Wu, Z.; Chen, C.; et al. 2024. NTIRE 2024 Image Shadow Removal Challenge Report. In *CVPRW*.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. In *Proceedings of the ACM Conference on Multimedia*.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2020. Underexposed Photo Enhancement using Deep Illumination Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, S.; Zheng, J.; Hu, H.-M.; and Li, B. 2013. Naturalness preserved enhancement algorithm for non-uniform illumination images. In *IEEE Transactions on Image Processing*.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*.

Wang, Y.; Wan, R.; Yang, W.; Li, H.; Lap-Pui, C.; and Kot, A. 2022. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. In *Proceedings of the Conference on British Machine Vision Conference*.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liang Liao, A. W.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; Geng Xue, W. S.; Yan, Q.; and Lin, W. 2024. Q-Instruct: Improving Low-level Visual Abilities for Multi-modality Foundation Models. In *CVPR*.

Wu, W.; Weng, J.; Zhang, P.; Wang, X.; Yang, W.; and Jiang, J. 2022. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *CVPR*.

Wu, Y.; Pan, C.; Wang, G.; Yang, Y.; Wei, J.; Li, C.; and Shen, H. T. 2023. Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement. In *CVPR*.

Xu, D.; Dong, W.; and Zhou, H. 2022. Sclera Recognition Based on Efficient Sclera Segmentation and Significant Vessel Matching. In *The Computer Journal*.

Xu, X.; Wang, R.; Fu, C.-W.; and Jia, J. 2022. Snr-aware low-light image enhancement. In *CVPR*.

Xu, X.; Wang, R.; and Lu, J. 2023. Low-Light Image Enhancement via Structure Modeling and Guidance. In *CVPR*.

Yan, Z.; Zhang, H.; Wang, B.; Paris, S.; and Yizhou, Y. 2016. Automatic Photo Adjustment Using Deep Neural Networks. In *ACM Transactions on Graphics*.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2020. From Fidelity to Perceptual Quality: A Semi-Supervised Approach for Low-Light Image Enhancement. In *CVPR*.

Yang, W.; Wang, W.; Huang, H.; Wang, S.; and Liu, J. 2021. Sparse gradient regularized deep retinex network for robust low-light image enhancement. In *IEEE Transactions on Image Processing*.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2023. Diff-Retinex: Rethinking Low-light Image Enhancement with A Generative Diffusion Model. In *ICCV*.

Yin, X.; Liu, X.; and Liu, H. 2019. FMSNet: Underwater Image Restoration by Learning from a Synthesized Dataset. In *International Conference on Artificial Neural Networks (ICANN)*.

Yin, Y.; Xu, D.; Tan, C.; Liu, P.; Zhao, Y.; and Wei, Y. 2023. CLE Diffusion: Controllable Light Enhancement Diffusion Model. In *Proceedings of the ACM Conference on Multimedia*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; ; and Shao, L. 2020. Learning enriched features for real image restoration and enhancement. In *ECCV*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, Y.; Zhang, J.; Guo, X.; and et al. 2019. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the ACM Conference on Multimedia*.

Zheng, N.; Zhou, M.; Dong, Y.; Rui, X.; Li, J. H. C.; and Zhao, F. 2023. Empowering lowlight image enhancer through customized learnable priors. In *ICCV*.

Zhou, D.; Yang, Z.; Yang, Y.; and et al. 2023a. Pyramid Diffusion Models For Low-light Image Enhancement. In *IJCAI*.

Zhou, H.; Dong, W.; Liu, X.; Liu, S.; Min, X.; Zhai, G.; and Chen, J. 2024. Low Light Image Enhancement via Generative Latent Feature based Codebook Retrieval. In *ECCV*.

Zhou, H.; Dong, W.; Liu, Y.; and Chen, J. 2023b. Breaking Through the Haze: An Advanced Non-Homogeneous Dehazing Method based on Fast Fourier Convolution and ConvNeXt. In *CVPRW*.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2020. EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*.