Zero-Reference Lighting Estimation Diffusion Model for Low-Light Image Enhancement

Jinhong He HEJH@STU.CQUT.EDU.CN
Minglong Xue✉ XUEML@CQUT.EDU.CN
Aoxiang Ning NINGAX@STU.CQUT.EDU.CN
Chengyun Song SCYER123@163.COM
Chongqing University of Technology

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Diffusion model-based low-light image enhancement methods rely heavily on paired training data, which limits its extensive application. Meanwhile, existing unsupervised methods lack effective bridging capabilities for unknown degradation. To address these limitations, we firstly propose a novel zero-reference lighting estimation diffusion model for low-light image enhancement called Zero-LED. It utilizes the stable convergence ability of diffusion models to bridge the gap between low-light domains and real normal-light domains and successfully alleviates the dependence on pairwise training data via zero-reference learning. Specifically, we first design the initial optimization network to preprocess the input image and implement bidirectional constraints between the diffusion model and the initial optimization network through multiple objective functions. Subsequently, the degradation factors of the real-world scene are optimized iteratively to achieve effective light enhancement. In addition, we explore a frequency-domain based and semantically guided appearance reconstruction module that encourages feature alignment of the recovered image at a fine-grained level and satisfies subjective expectations. Finally, extensive experiments demonstrate the superiority of our approach to other state-of-the-art methods and more significant generalization capabilities.

Keywords: Low-light Image Enhancement, Zero Reference Learning, Diffusion Models, Multi-modal

1. Introduction

Low light enhancement aims to enhance the quality and brightness of under-illuminated images. Due to the complex lighting conditions in the real world, relevant information in captured images is often lost through appropriate or significant masking. This poses a challenge to human visual perception and impedes the development and deployment of various downstream tasks (e.g., Autonomous Driving Li et al. (2021b), Target Detection Liang et al. (2021), Text Detection Xue et al. (2020)). Low-light image enhancement techniques have been significantly developed to address these challenges recently, with many related algorithms proposed. These techniques can be broadly categorized into traditional model-based approaches and data-driven deep learning-based approaches. The former primarily involves constructing physical models through methods such as histogram equalization Pisano et al. (1998) and Retinex theory Land and McCann (1971), which are then processed through manual a priori optimization of model parameters and the information inherent in the image itself Park et al. (2022); Fu et al. (2016). The effectiveness of these traditional methods

© 2024 J. He, M. Xue✉, A. Ning & C. Song.

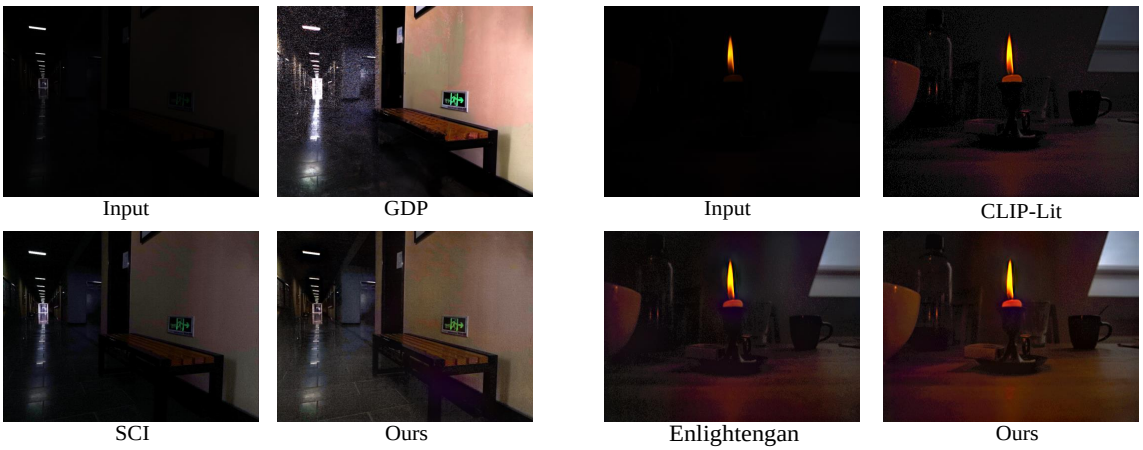

Figure 1: Comparison between state-of-the-art unsupervised methods and our method. It can be seen that these comparison methods appear to suffer from excessive noise, color distortion, and visual quality degradation.

relies heavily on the accuracy of the manual prior assumptions. However, lighting conditions are inherently ill-defined in the real world, leading to difficulties elucidating various low-light factors.

With the development of deep learning, the quest to extract a priori knowledge from massive datasets has given rise to numerous data-driven network learning methods (Sun et al. (2015); Jiang et al. (2023)). Supervised learning-based methods utilize paired low-light/high-light image data to establish the corresponding association mappings between low-light and clear images for learning directly, which is more robust and effective than traditional methods. Despite these advances, there are still significant challenges in constructing paired training data, which has led to the exploration of low-light enhancement methods based on unsupervised learning Guo et al. (2020); Jiang et al. (2021); Liang et al. (2023). Essentially, they all work by bridging the mapping between the input and output domains to obtain clear highlighted images. However, the existing unsupervised methods make it challenging to generate a reconstruction of the content structure due to the lack of effective supervision. They can also not effectively generate and optimize the unknown degradation factors of realistic complex scenes, which often cause excessive noise and artifacts to the extent that it is difficult to obtain satisfactory visual restoration, as shown in Fig. 1. In particular, numerous low-light image enhancement algorithms based on the diffusion model (Jiang et al. (2023); Fei et al. (2023)) have been proposed recently and achieved strong performance results. However, due to the stochastic nature of the diffusion inference process and the dependence on supervisory constraints, most of these algorithms are based on supervised training with paired datasets or optimizing the network using a priori knowledge, which makes it challenging to achieve genuinely effective unsupervised diffusion training and widely deployed in real-world.

Therefore, to alleviate the above problems, we explore a bidirectionally constrained unsupervised diffusion training approach to address the lack of generative power during unsupervised training and the unknown degradation of realistic low-light scenes. Specifically, we propose a light estimation-based diffusion model for zero-reference low-light image

enhancement called Zero-LED. A pluggable initial optimization network is constructed via a deep network Guo et al. (2020) for the preprocessing of diffusion, which is passed as a structural constraint to the diffusion process to mitigate the stochastic nature of diffusion. Subsequently, the light estimation of the inference process is optimized by an objective function, and this is passed in reverse to optimize the initial optimization network to achieve iterative bidirectional supervisory optimization. In addition, to reduce the consumption of computational resources, we transfer the diffusion process to be carried out on the wavelet domain by wavelet transform. we design a text- and frequency-domain based Appearance Reconstruction Module (ARM), which guides the inference output of diffusion through different modalities and combines the efficient capture of detailed content in the frequency-domain space to achieve metrics-favorable and perceptually oriented enhancement effects. Benefiting from these, our approach has a more significant generalization capability to the real world, and extensive experiments on publicly available datasets demonstrate the superiority of our approach over other state-of-the-art unsupervised methods. Overall, our contributions are as follows:

- To our best knowledge, we firstly propose a zero-reference low-light image enhancement diffusion model that effectively enhances low-light images through a bidirectional optimization approach, thus reducing the dependence on paired training data. The model has enhanced the generative ability to bridge the gap between normal and low-light domains, introducing new perspectives for future research.
- We design a semantic and frequency domain-based appearance reconstruction module. It utilizes different modalities and multiple frequency domain spaces to constrain the stochastic nature of the diffusion inference process and efficiently reconstructs images for better perceptual results.
- Extensive experiments on real-world based datasets have demonstrated the superiority of our method over other state-of-the-art methods, as well as more significant generalization capabilities.

2. Related Work

2.1. Low-light Image Enhancement

Early techniques for enhancing low-light images primarily focused on employing a model-based approach to optimize the parametric information within the image itself for processing. The Retinex theory Land and McCann (1971) achieves the desired reflectance map (standard image) by eliminating the low-light input's illumination. However, these model-based approaches often depend on manual crafting of many a priori assumptions and struggle to adapt to the complex environmental conditions of natural scenes.

Therefore, along with the proven benefits of deep learning in numerous low-level vision tasks, researchers have also focused their attention on low-light image enhancement algorithms (Wei et al. (2018); Wu et al. (2022); Xue et al. (2024)) that leverage a data-driven approach for efficient generalization through deep learning. For example, Chen et al. Chen et al. (2018) curated a dataset containing different exposure levels for nighttime imaging challenges. Wei et al. Wei et al. (2018) designed an end-to-end trainable RetinexNet. However, due to the challenges of acquiring paired low-light image datasets, Jiang Jiang et al.

(2021) utilized generative adversarial Networks (GANs) as the main framework, pioneering unpaired images for training. Similarly, Guo et al. Guo et al. (2020) elaborated a pixel-level curve estimation convolutional neural network by iterative derivation to establish a reference-free training paradigm. Liang et al. Liang et al. (2021) designed a Retinex architecture-based search unfolding technique. Yang et al. Yang et al. (2023) employed neural representation to normalize the degradation to alleviate the enhancement difficulty. Despite the achievements of these unsupervised methods, their generalization to real-world scenarios is still limited. For this reason, we adapt to various complex environmental conditions by diffusing the generative power of the model.

2.2. Diffusion Model for Image Restoration

Recently, diffusion models Song and Ermon (2019) have garnered significant acclaim within image generation by leveraging parametric Markov chains to optimize the lower variational bounds of the likelihood function. This enables them to yield more precise target distributions than alternative generative models, such as GANs. Concurrently, to amplify the generative prowess of algorithms in image restoration, many researchers have embarked on developing various restoration endeavours grounded in diffusion models. Saharia et al. Saharia et al. (2022) adopt a direct cascading approach, integrating low-resolution measurements and latent codes as inputs to train conditional diffusion models for restoration. Jiang et al. Jiang et al. (2023) advances a diffusion model rooted in wavelet transform tailored for enhancing images captured in low-light environments, achieving content stabilization through forward diffusion and denoising processes during training. WeatherDiff Özdenizci and Legenstein (2023) introduces a block-based diffusion model aimed at recuperating images taken in adverse weather conditions, employing guidance across overlapping blocks during the inference stage. Additionally, Fei et al. Fei et al. (2023) utilize the a priori knowledge embedded in a pre-trained diffusion model to effectively address any linear inverse problem.

Although diffusion models have achieved satisfactory visual restoration, due to the uncontrollable nature of diffusion, these algorithms are almost always based on supervised training on paired datasets or network optimization using the a priori knowledge of pre-trained diffusion models. It is still a great challenge to realize unsupervised training. In this paper, we propose a bi-directionally constrained unsupervised diffusion training approach to achieve robust zero-reference trained diffusion models for the first time, as well as more significant generalization ability and effective low-light image enhancement.

3. Methodology

The main goal of this paper is to explore a diffusion model based on zero-reference training, and the overall framework is shown in Fig. 2. Leveraging the generative capacity of the diffusion model, the proposed method achieves notable enhancements in image quality. By developing a bidirectional optimization training method, we establish a diffusion model based on zero-reference images, thereby reducing reliance on training data and enhancing generalizability to real-world contexts. Furthermore, to minimize computational resource consumption and enhance efficiency Jiang et al. (2023), we transition the diffusion inference process to the wavelet low-frequency domain via wavelet transformation. In this

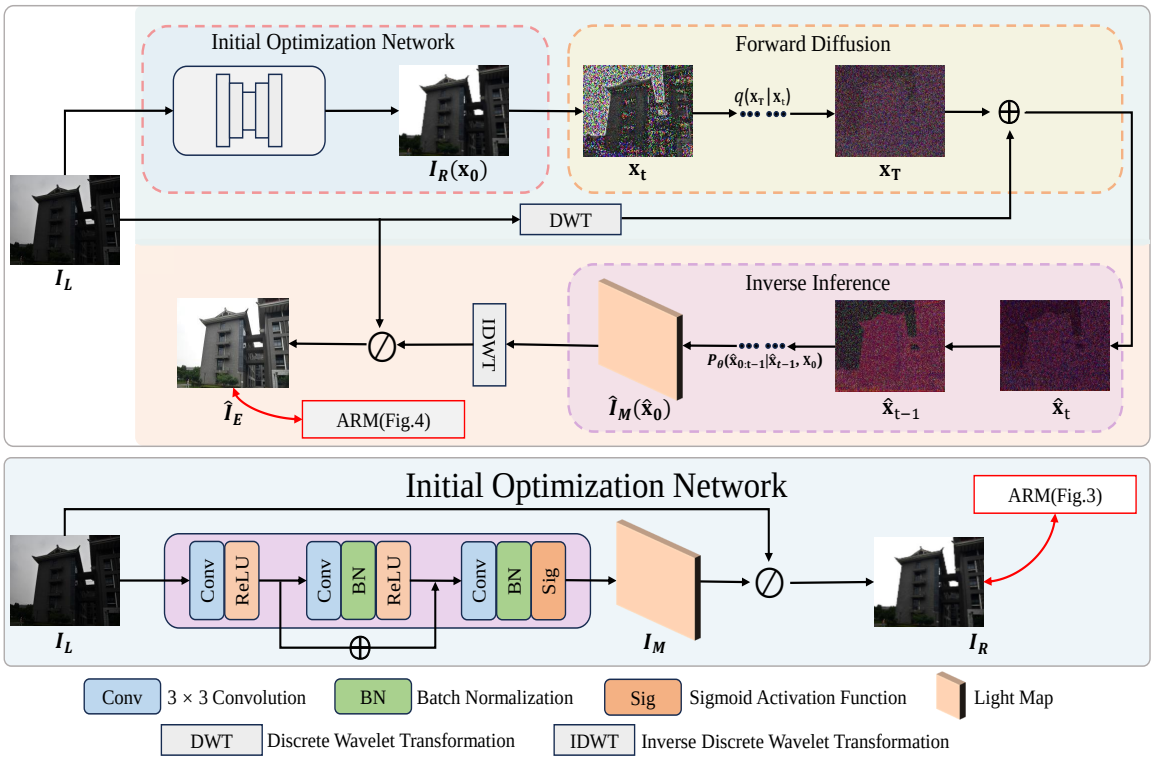

Figure 2: The overall framework of Zero-LED. It proposes a Bidirectional optimization approach combining a deep neural network and a diffusion model for training without reference images. The initial optimization network provides the structural image and preliminary optimization of unknown degradation factors for the diffusion process. The inference process further bridges the gap between degraded and normal light and is optimized by an objective function in both directions.

section, we provide a detailed exposition of the underlying principles of the traditional conditional diffusion model and the crux of our proposed methodology, namely, the bidirectional optimization-based zero-reference diffusion model. Lastly, we introduce the Appearance Reconstruction Module (ARM), grounded in text and frequency domains, as a meticulously crafted component adept at guiding the reconstruction of image content structure and the overarching enhancement of quality.

3.1. Conditional Diffusion Models

Diffusion models Ho et al. (2020); Song et al. (2020) to train Markov chains by variational inference. It converts complex data into completely random data by adding noise and gradually predicts the noise to recover the expected clean image. Consequently, it usually includes the forward diffusion process and reverse inference process.

Forward Diffusion Process. The forward diffusion process primarily relies on incrementally introducing Gaussian noise with a fixed variance $\{\beta_t \in (0, I)\}_{t=1}^{T}$ into the input distribution x_0 until the T time steps approximate purely noisy data. This process can be

expressed as:

$$q(x_1, \cdots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{2}$$

where x_t and β_t are the corrupted noise data and the predefined variance at time step t. Respectively, N denotes a Gaussian distribution. Furthermore, each time step x_t of the forward diffusion process can be obtained directly by calculating the following equation, computed from x_0:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad \epsilon \sim N(0, I), \tag{3}$$

where $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t$=$\prod_{i=1}^{t} \alpha_i$.

Reverse Inference Process. The reverse inference process aims to restore the original data from the generated Gaussian noise. In contrast to the forward diffusion process, where the data distribution at each time step t can be directly computed using a formula, the reverse process relies exclusively on iteration to eliminate the predicted noise and restore the data until the randomly sampled noise $x_T \sim N(0, I)$ into the clean data $\hat{x}_0$. Formulated as:

$$p_\theta(\hat{x}_0, \cdots, \hat{x}_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(\hat{x}_{t-1}|\hat{x}_t), \tag{4}$$

$$p_\theta(\hat{x}_{t-1}|\hat{x}_t) = N(\hat{x}_{t-1}; \mu_\theta(\hat{x}_t, t), \sigma_t^2 I), \tag{5}$$

where μ_θ is the diffusion model noise predictor, which is mainly optimized by the editing and data synthesis functions and used as a way to learn the conditional denoising process, as follows:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)), \tag{6}$$

where ϵ_θ is a function approximator intended to predict ϵ from x_t.

3.2. Diffusion Models for Zero-reference Learning

Existing conditional diffusion models have achieved significant performance, but the substantial demands on computational resources and paired datasets hinder their practical deployment in image restoration tasks. To address these challenges, we propose a zero-reference learning approach for diffusion models, which includes the following components:

Initial Optimization Network. To consider the generative diversity of the diffusion model, we use an initial optimization network as the a priori network for diffusion training. Specifically, we do this by preprocessing the input image and passing it into the diffusion process. Compared with the input image, the preprocessed structural image I_R has clearer image content. The generation of chaotic content can be constrained more precisely in the inference process. In addition, the initial optimization network provides an initial degradation domain calibration for the diffusion model and combines the objective function to form an iterative optimization with the diffusion model.

Based on a kind of consensus Ma et al. (2022); Guo et al. (2020) that there is a link between illumination and low-light images, as well as the consideration of model complexity. Therefore, we plan to decompose the low light image I_L using a lightweight deep neural network to obtain its illumination component $I_M \in R^{H\times W\times C}$. Additionally, based on the Retinex theory (i.e. $I_L = I_R \otimes I_M$), we can preliminarily obtain the structural image I_R.

Diffusion-based Degradation Model. Inspired by Jiang et al. Jiang et al. (2023), we utilize the wavelet transform to process the input image. The use of the discrete wavelet transforms ($DWT(\cdot)$) enables the extraction of the low-frequency space of the low-light image, thereby reducing the computational resources required by the diffusion model. The input image is decomposed into its low-frequency information (L) and high-frequency information (H) through the $DWT(\cdot)$. Furthermore, the high-frequency space (H) produced by the wavelet transform includes three subbands in the vertical, horizontal, and diagonal directions. As a result, the low-frequency information that has been decomposed (L) can be considered a version of the image that has been reduced by a quarter dimension, significantly reducing the demand for computer resources.

In the real world, images are often subject to unknown degradation factors in complex, dimly lit environments, and thus, low-light image enhancement is often seen as a task to construct unknown degradation models. We further simulate complex degradation based on the structured images I_R provided by the initial optimization network and combined with the generative power of the diffusion model. The diffusion model is also used to generate fine-tun illumination masks to achieve significant enhancement effects. This can be described as:

$$\hat{I}_E = I_L \oslash \mathcal{D}(I_R, I_L), \tag{7}$$

where $\mathcal{D}$ is a diffusion-based degenerate model. Specifically, the inference process Eq. 4 $p_\theta(\hat{x}_0, \cdots, \hat{x}_{T-1}|x_T)$ will be carried out under the structural image I_R and low light image I_L. Our goal is to learn the degradation parameters of the image from the denoising process $p_\theta(\hat{x}_{0:T}|(I_R, I_L))$ while guaranteeing high fidelity of the sampling results to its generated content. Additionally, we will optimize the noise predictor to fit the illumination component $\hat{I}_M$ and minimize the L2 distance between $\hat{I}_M$ and I_M to refine its degeneracy parameters and optimize the initial optimization network. Therefore, the objective function of the optimized diffusion model can be expressed as follows:

$$\mathcal{L}_{diff} = E_{t\sim[1,T]} E_{x_0\sim p(x_0)} E_{\epsilon_t\sim N(0,I)} \parallel \epsilon_t - \epsilon_\theta(x_t, t) \parallel^2 + ||\hat{I}_M - I_M||^2. \tag{8}$$

In addition, smoothing loss with a spatially varying number of paradigms Ma et al. (2022); Fan et al. (2018) is used to optimize the predicted illumination component, denoted as:

$$\mathcal{L}_{smooth} = \sum_{k=1}^{K} \sum_{n\in K(k)} \gamma_{k,n}(\parallel \hat{I}_{M(k)} - \hat{I}_{M(n)} \parallel_1 + \parallel I_{M(k)} - I_{M(n)} \parallel_1), \tag{9}$$

Where K is the total number of pixels, k is the kth pixel. $\gamma_{k,n}$ denotes the weight.

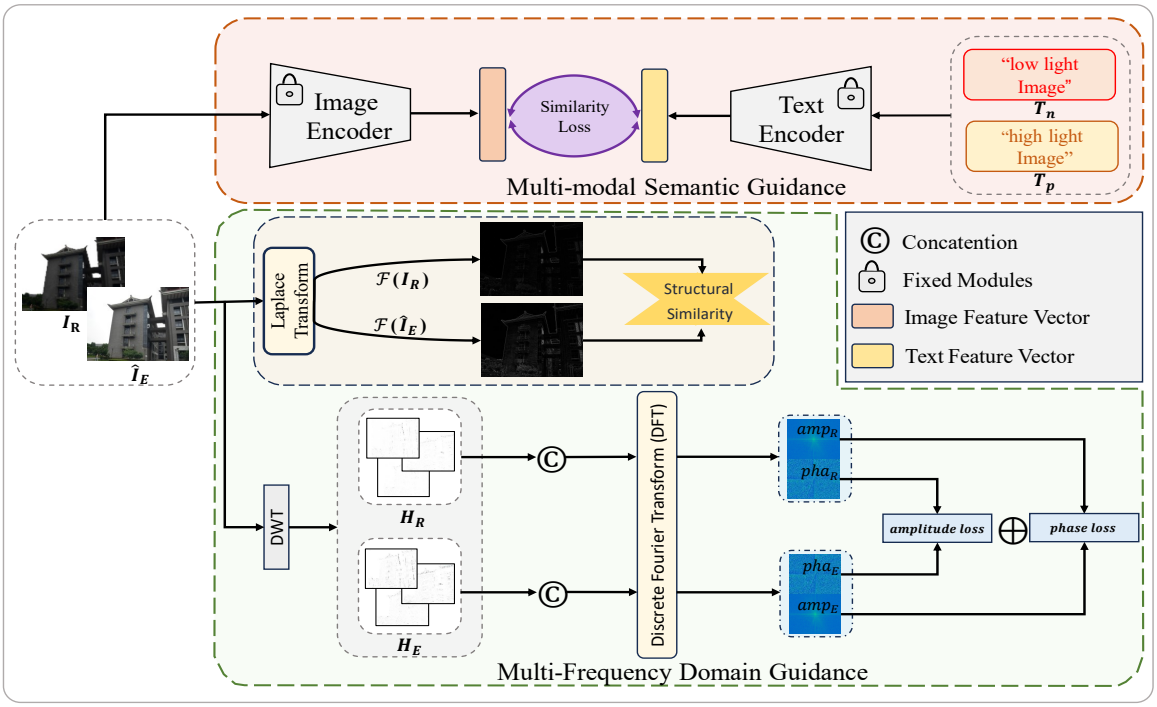

Figure 3: Framework diagram of our proposed appearance reconstruction module. multi-modal semantics focuses on guiding illumination enhancement and supervising the input of image and text features. Frequency-domain guidance focuses on supervising high-frequency details and constraining the generation of artifacts.

3.3. Appearance Reconstruction Module

Existing unsupervised methods hardly obtain significant detail features from low-light images for satisfactory restoration, while diffusion models are complex to perform effective content structure recovery in unsupervised training scenarios. Therefore, as shown in Fig. 3, to achieve good reconstruction of the generated content and perceptually oriented enhancement, we propose a semantic and multi-frequency domain guided appearance reconstruction module based on semantic and multi-frequency domain guidance to obtain efficient appearance reconstruction.

Multi-modal Semantic Guidance. Recent studies have shown Liang et al. (2023); Yang et al. (2023) that multi-modal learning can lead to the effective enhancement of low-light images compared to a single modality. Therefore, we use a pre-trained CLIP model to acquire prior knowledge. Precisely, we extract the latent codes of the cue pairs by feeding predefined cue pairs containing positive prompts T_p and negative prompts T_n (as shown in Fig. 3) to the text encoder (Φ_{text}). Correspondingly, the latent codes of the prediction result $\hat{I}_E$ and the structure image I_R are extracted by the image encoder (Φ_{image}). We then measure the difference between image vectors and text vectors by computing the similarity loss between them in the CLIP latent space:

$$\mathcal{L}_{clip} = \sum_{I \in \{I_E, \hat{I}_E\}} \frac{e^{\cos(\Phi_{image}(I), \Phi_{text}(T_n))}}{\sum_{T \in \{T_p, T_n\}} e^{\cos(\Phi_{image}(I), \Phi_{text}(T))}}. \tag{10}$$

In addition, we present the hyper-parameters υ as the probability value of normal light to incentivize the prediction results further to achieve good illumination. Semantically consistent output is encouraged by minimizing the distance between the positive prompts probability and the hyper-parameters υ:

$$\mathcal{L}_{prob} = \| \, cos(\Phi_{image}(\hat{I}_E), \Phi_{text}(T_n)) - \upsilon \|_1. \tag{11}$$

Frequency Domain Guidance. In this study, we combine the advantages of spatial information capture in multiple frequency domains. The spectrum is utilized to help diffusion models perform appearance reconstruction during unsupervised training, leading to metric-friendly and perception-oriented enhancements.

We first perform the Laplace transform $\mathcal{F}(\cdot)$ on the structural image I_R to supervise the sampling results from the edge level. Simultaneously, we implemented constraints on the content of the image generation using SSIM content loss Wang et al. (2004) with an initial optimization network:

$$\mathcal{L}_{content} = (1 - SSIM(\mathcal{F}(\hat{I}_E), \mathcal{F}(I_R))) + (1 - SSIM(\hat{I}_E, I_R)). \tag{12}$$

In addition, we combine wavelet transform and Fourier transform to capture the deep content features of the image and reconstruct the image using spectrum. Compared to the wavelet low-frequency domain, the wavelet high-frequency domain contains only the structural details of the image and is not prone to content loss. Therefore, we perform a discrete Fourier transform $DFT(\cdot)$ in the wavelet high-frequency domain of the predicted image $\hat{I}_E$ and the structural image I_R to obtain their corresponding amplitude and phase (amp, pha):

$$amp_E, pha_E = DFT(DWT(\hat{I}_E)), \tag{13}$$

$$amp_R, pha_R = DFT(DWT(I_R)). \tag{14}$$

To encourage the recovery results to be feature-aligned at a fine-grained level and cross-validated with the optimization network. For this purpose, we use L1 loss to minimize the information differences between spectrograms:

$$\mathcal{L}_{spectral} = \vartheta_1 \parallel amp_E - amp_R \parallel_1 + \vartheta_2 \parallel pha_E - pha_R \parallel_1, \tag{15}$$

where ϑ_1 and ϑ_2 are weighting parameters for amplitude and phase losses. Thus, for multi-modal semantics and frequency domain-guided appearance reconstruction, the total loss can be summarised as:

$$\mathcal{L}_{rec} = \mathcal{L}_{content} + \mathcal{L}_{spectral} + \varpi(\mathcal{L}_{prob} + \mathcal{L}_{clip}), \tag{16}$$

where ϖ is the weight of the semantic guidance loss.

3.4. Network Training

Besides the objective functions used to optimise the diffusion model and appearance reconstruction, we also utilise two quality-enhancing losses and the MSE to improve the quality of the final output and extend the network learning capabilities.

Color Constancy Loss. Based on the grey world color constancy hypothesis Buchsbaum (1980). We designed a color constancy loss to correct for potential color bias in the enhanced image and also established a relationship between the three adjustment channels:

$$\mathcal{L}_{col} = \sum_{\forall(m,n)\in\varrho} (C^m - C_n)^2, \varrho = \{(R,G),(R,B),(G,B)\}, \tag{17}$$

where C^m is the average intensity value of m channel in the recovered image, (m,n) represents a pair of channels.

Spatial Consistency Loss. The loss of spatial coherence constraints the differences in neighbouring regions between the input image and the enhanced image:

$$\mathcal{L}_{spa} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\Theta(i)} (\parallel E_i - E_j \parallel_1 - \parallel I_i - I_j \parallel_1)^2, \tag{18}$$

where N is the number of local regions and $\Theta(i)$ is the four neighbouring regions (top, bottom, left, and right) centred on region i. We denote E and I as the average intensity values of the local regions in the enhanced version and the low-light image, respectively.

The total loss $\mathcal{L}_{total}$ is expressed by combing the diffusion objective function, the appearance reconstruction loss, and the Quality Enhancement Losses as:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \omega\mathcal{L}_{smooth} + \mathcal{L}_{rec} + \mathcal{L}_{col} + \mathcal{L}_{spa}, \tag{19}$$

where ω is the weight of the illumination smoothing loss.

4. Experiments

4.1. Experimental Settings

Implementation Details. We implemented our method using Pytorch on two NVIDIA RTX 3090 GPUs. We set the total number of training iterations to 5×10^4, using the Adam optimizer with the initial learning rate set to 1×10^{-4}. The batch size and patch size were set to 4 and 256×256, respectively.

Benchmark Datasets. Our network is trained and evaluated to validate the effectiveness of our method on the LSRW Hai et al. (2023) dataset, in which we randomly select 1000 low-light-normal-light image pairs for training and 50 image pairs for evaluation. Most low-light images were collected realistically by varying the exposure time and ISO and fixing other camera parameters. In addition, we extend several real-world benchmark datasets to evaluate the performance of our proposed network to increase persuasiveness. Examples include LOLv1 Wei et al. (2018), LIME Guo et al. (2016) and Backlit300 Liang et al. (2023). The dataset of LOLv1 contains 500 real-world low/normal light image pairs, of which 485 image pairs are used for training and 15 image pairs are used for evaluation. Also, to

Methods	Reference	LOLv1				LSRW				LIME		Backlit300	
		PSNR↑	SSIM↑	NIQE↓	LOE↓	PSNR↑	SSIM↑	NIQE↓	LOE↓	NIQE↓	LOE↓	NIQE↓	LOE↓
Zero-DCE	CVPR'20	14.861	0.559	11.985	215.816	15.801	0.446	11.832	247.291	11.942	192.089	16.026	165.325
Zero-DCE++	TPAMI'21	14.682	0.472	10.646	277.736	15.791	0.457	11.341	241.348	11.376	296.654	14.693	285.720
Enlightengan	TIP'21	17.606	0.653	9.996	365.561	17.136	0.460	11.937	385.135	14.585	421.018	15.058	385.796
RUAS	CVPR'21	16.405	0.499	10.725	125.351	14.271	0.411	11.081	198.930	12.413	288.730	14.486	598.305
SCI	CVPR'22	14.784	0.521	11.827	101.113	15.241	0.419	10.774	234.605	12.379	212.621	13.376	298.768
CLIP-Lit	ICCV'23	12.394	0.493	12.187	355.441	13.483	0.405	9.144	289.583	12.239	192.001	16.633	195.875
GDP	CVPR'23	15.896	0.542	10.273	120.278	12.887	0.362	9.178	75.884	13.138	78.929	13.693	148.929
Ours	-	16.632	0.566	8.355	148.563	15.824	0.461	8.381	175.355	10.843	146.663	11.993	351.877

Table 1: Quantitative evaluation of different unsupervised learning methods on four benchmark datasets. The best and second performance are marked in red and blue, respectively.

Figure 4: Visual comparison of low-light enhancement methods on the LSRW dataset.

demonstrate the generalization to real-world degraded scenes, we evaluate the generalization ability of the proposed method in this paper by the benchmark dataset LIME and by selecting 30 degraded images in the Backlit300 dataset. Note that during the training process, we only use low-light images from the paired dataset and do not use normal-light images to demonstrate the superiority of our zero-reference method.

Metrics. We propose to evaluate the performance of different algorithms using two full-reference distortion metrics, PNSR and SSIM Wang et al. (2004), and two non-reference metrics, NIQE Mittal et al. (2012) and LOE Wang et al. (2013). Higher PSNR or SSIM implies more realistic restoration results, while lower NIQE or LOE indicates higher quality details, luminance, and hue.

4.2. Comparison with the State-of-the-Art

Comparison Methods. To verify the effectiveness of the method proposed in this paper, we compare it with the state-of-the-art unsupervised learning methods in recent years, i.e., Zero-DCE Guo et al. (2020), Zero-DCE++ Li et al. (2021a), RUAS Liu et al. (2021), Enlightengan Jiang et al. (2021), SCI Ma et al. (2022), CLIP-Lit Liang et al. (2023) and GDP Fei et al. (2023).

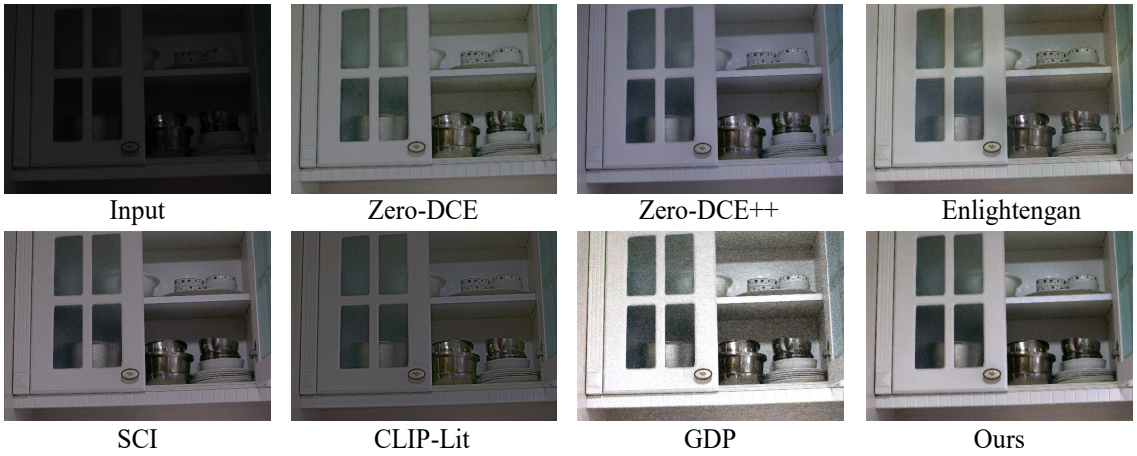

Figure 5: Visual comparison of low-light enhancement methods on the LOLv1 dataset.

Quantitative Comparison. We obtained quantitative results for the other methods using official pre-trained models and running their respective public codes. As shown in Table. 1, our method achieves quantitative performance close to the state-of-the-art on several metrics compared to all compared methods. Among them, we obtain the second-best results in the LOLv1 dataset regarding the full-reference distortion metric PSNR/SSIM. In the LSRW dataset, we obtained the second-best results for the PSNR metric. In addition, for the no-reference metrics NIQE/LOE, our method obtains the lowest NIQE scores on all datasets and the second-best LOE metric evaluation results on both datasets. Our method can better balance the quantitative results of images. This fully demonstrates the superiority of our method and its better generalization ability in real-world scenarios.

Qualitative Comparison. For a more intuitive comparison, we report the visual results of all methods in Fig. 4. By visual comparison, our methods achieve visually pleasing results with improved color and brightness. In contrast, previous state-of-the-art unsupervised learning methods produce artifacts and unnatural tones by producing excessive smoothing or struggling to adapt effectively to degradation factors due to a lack of practical constraints and guidance. For example, Enlightengan Jiang et al. (2021) produces artifacts, and CLIP-Lit Liang et al. (2023) produces excessive color effects. In particular, Diffusion prior-based GDP Fei et al. (2023) hardly enhances low-light images extensively. Furthermore, we visually compare the LOLv1 test set in Fig. 5. In contrast, other methods fail to recover enough brightness (SCI Ma et al. (2022)) or generate too much noise (GDP Fei et al. (2023)). Our approach provides more substantial constraints, including visually orientated guidance from prompts, thereby producing a more natural visual perception.

4.3. Ablation Studys

The effectiveness of bidirectional optimization training. To validate the importance of the initial optimization network, we consider two training approaches for ablation studies. We used the LOLv1 dataset for all ablation experiments. "#1" indicates that the initial optimization network was removed from the overall architecture and trained by directly inputting low-light images for diffusion enhancement. Meanwhile, we used the full Zero-

Figure 6: Qualitative results from ablation studies. Networks with complete structures perform best.

index	PSNR↑	SSIM↑	NIQE↓	LOE↓
#1	12.023	0.413	12.579	132.845
ARM_1	-	-	-	-
ARM_2	13.290	0.498	11.395	141.526
ARM_3	14.559	0.512	10.613	137.836
Zero-LED	16.432	0.562	8.355	148.563

Table 2: Quantitative evaluation of the enhancement results obtained from different settings. Results using the complete settings are underlined.

LED for two-way optimization training. Firstly, as shown in Table. 2, we evaluate the results of #1, whose enhancement effect is very different from the final results, which illustrates the importance of the initial supervised network. Moreover, as shown in Fig. 6, we can intuitively conclude that Zero-LED produces clearer results and better perceptual effects than "#1". This fully validates the iterative optimization effect of our two-way optimization training method.

The effectiveness of appearance reconstruction based on multi-modal semantics and frequency domain space. We divided the appearance reconstruction module into three versions by incrementally adding module sections. In the ARM_1 version, we remove the appearance reconstruction module and use only image-level supervision. In the ARM_2 version, we reconstruct the image using a Fourier transform using the wavelet high-frequency domain. In the ARM_3 version, we reconstruct the image using the default frequency domain space of the paper. Finally, we used the full Zero-LED, added multimodal semantics, and combined it with the frequency domain space to restore the appearance of the image. As shown in Figure 6, in the ARM_1 version, we were unable to reconstruct the image efficiently. With the addition of the frequency domain space, the ARM_2 and ARM_3 versions restore objects in dark areas to a large extent, but the enhanced brightness and colors are still unrealistic. On the other hand, Zero-LED results in more realistic brightness and the best-perceived effect. This demonstrates the importance of the frequency domain space for image content reconstruction and the effectiveness of multimodal semantics in guiding image appearance. In addition, as shown in Table 2, ARM_2 shows a significant decrease in metrics evaluation compared to ARM_3, but the best performing quantitative result is still Zero-LED.We attribute this to the fact that the wavelet high-frequency domain leads to more content loss when performing the Fourier Transform and the text-guided appearance being visually friendly for Zero-LED.

5. Conclusion

We firstly propose a bidirectional zero-reference training approach via an initial optimization network and successfully implement a zero-reference trained diffusion model called Zero-LED. First, for the input degraded images, we perform preliminary fitting of the degradation parameters via an initial optimization network and acquire the structural images. We obtain a better lighting estimation with the calibration based on the diffusion model. In addition, we propose a text- and frequency-domain-based appearance reconstruction module for the output restored image, which provides perceptually oriented restoration guidance using a pre-trained visual language model and multiple frequency-domain spaces to guide the restoration of structural content jointly. Experimental results on publicly available benchmark tests show that our approach outperforms competitors in the comprehensive evaluation while providing better stability and generalization.

Acknowledgments

This work was supported in part by the Chongqing Postgraduate Research and Innovation Project Funding (Grant No. CYS240680), Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202401106).

References

Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.

Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6306–6314, 2018.

Qingnan Fan, Jiaolong Yang, David Wipf, Baoquan Chen, and Xin Tong. Image smoothing via unsupervised learning. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018.

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9935–9946, 2023.

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.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020.

Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.

Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90:103712, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971.

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

Guofa Li, Yifan Yang, Xingda Qu, Dongpu Cao, and Keqiang Li. A deep learning based image enhancement approach for autonomous driving at night. *Knowledge-Based Systems*, 213:106617, 2021b.

Jinxiu Liang, Jingwen Wang, Yuhui Quan, Tianyi Chen, Jiaying Liu, Haibin Ling, and Yong Xu. Recurrent exposure generation for low-light face detection. *IEEE Transactions on Multimedia*, 24:1609–1621, 2021.

Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8094–8103, 2023.

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.

Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022.

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

Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Jaemin Park, An Gia Vien, Jin-Hwan Kim, and Chul Lee. Histogram-based transformation function estimation for low-light image enhancement. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1–5. IEEE, 2022.

Etta D Pisano, Shuquan Zong, Bradley M Hemminger, Marla DeLuca, R Eugene Johnston, Keith Muller, M Patricia Braeuning, and Stephen M Pizer. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital imaging*, 11:193–200, 1998.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 769–777, 2015.

Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013.

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.

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

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.

Minglong Xue, Palaiahnakote Shivakumara, Chao Zhang, Yao Xiao, Tong Lu, Umapada Pal, Daniel Lopresti, and Zhibo Yang. Arbitrarily-oriented text detection in low light natural scene images. *IEEE Transactions on Multimedia*, 23:2706–2720, 2020.

Minglong Xue, Jinhong He, Yanyi He, Zhipu Liu, Wenhai Wang, and Mingliang Zhou. Low-light image enhancement via clip-fourier guided wavelet diffusion. *arXiv preprint arXiv:2401.03788*, 2024.

Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12918–12927, 2023.