Low-light Image Enhancement via CLIP-Fourier Guided Wavelet Diffusion

Minglong Xue, Jinhong He, Wenhai Wang and Mingliang Zhou

***Abstract*—Low-light image enhancement techniques have significantly progressed, but unstable image quality recovery and unsatisfactory visual perception are still significant challenges. To solve these problems, we propose a novel and robust low-light image enhancement method via CLIP-Fourier Guided Wavelet Diffusion, abbreviated as CFWD. Specifically, CFWD leverages multimodal visual-language information in the frequency domain space created by multiple wavelet transforms to guide the enhancement process. Multi-scale supervision across different modalities facilitates the alignment of image features with semantic features during the wavelet diffusion process, effectively bridging the gap between degraded and normal domains. Moreover, to further promote the effective recovery of the image details, we combine the Fourier transform based on the wavelet transform and construct a Hybrid High Frequency Perception Module (HFPM) with a significant perception of the detailed features. This module avoids the diversity confusion of the wavelet diffusion process by guiding the fine-grained structure recovery of the enhancement results to achieve favourable metric and perceptually oriented enhancement. Extensive quantitative and qualitative experiments on publicly available real-world benchmarks show that our approach outperforms existing state-of-the-art methods, achieving significant progress in image quality and noise suppression. The project code is available at https://github.com/hejh8/CFWD.**

***Index Terms*—Low-light image enhancement, diffusion model, multi-modal, Fourier transform, wavelet transform.**

I. Introduction

LOW-Light image 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, such as Target Detection [1], Autonomous Driving [2] and Text Detection [3]. Therefore, to address these challenges, low-light image enhancement techniques have been vigorously developed, and many related algorithms have been proposed. These techniques can be broadly categorized into traditional model-based approaches and data-driven deep learning-based approaches.

Traditional model-based low illumination image enhancement methods mainly construct physical models through methods such as histogram equalization [4] and Retinex theory [5]. Their focus is on using manually designed prior knowledge [6]–[9] to optimize the degradation parameters of the image itself, and the effectiveness relies heavily on the accuracy of the manually created prior. However, low-illumination image enhancement is essentially a nonlinear problem with unknown degradation, so it is more difficult to use an artificial prior to adapt to various lighting conditions in an open scene.

With the development of deep learning, researchers have explored a large number of data-driven-based network learning methods [10]–[17]. Wei et al. [18] constructed a deep-learning image decomposition algorithm based on the Retinex model. xu et al. [19] utilized a signal-to-noise ratio-aware transformer and a convolutional neural network (CNN) with spatially varying operations for restoration. In addition, the recently emerged diffusion model [20], [21] has attracted extensive attention from researchers in the field of image restoration [22]–[24] due to its powerful generative and generalization capabilities. These methods essentially bridge the gap between the degraded and normal domains to obtain a clear normal image.

However, most existing methods such as GSAD and SNR-Net tend to consider only supervising the enhancement process from the image level, neglecting the detailed reconstruction of the image and the role of multi-modal semantics in guiding the feature space. Such unimodal supervision produces sub-optimal reconstruction of uncertain regions and poorer local structures, leading to the appearance of unsatisfactory visual results. For example, as shown in Fig. 1, previous state-of-the-art approaches can suffer from color distortion, excessive noise, and redundant confusing information due to the lack of effective constraints and guidance. It is worth noting that diffusion models have diverse generative effects due to the stochastic nature of the inference process but also indirectly contribute to the difficulty of efficiently constraining noise and redundant information in image restoration tasks.

Furthermore, for low-level visual tasks, the simple introduction of visual-language information does not reap significant performance. This may be due to the fact that image corruption creates difficulties for feature alignment, resulting in the inability of the visual-language model to capture the fine-grained gaps between degraded images and semantics effectively. Considering the above issues, our overall goal is to explore the introduction of multimodal semantics through frequency-domain diffusion iterations based on the Contrast-Language-Image-Pre-Training (CLIP) model to provide effective condition guidance and content constraints for the task of low-light image enhancement, and to achieve the enhancement of low-light image under different spatial illumination conditions.

Inspired by [25], we adopt the wavelet diffusion model to establish a mapping between low-light and normal-light images, and also propose a novel CLIP and Fourier transform guided wavelet diffusion model (CFWD). Specifically, based on the pre-trained visual-language model CLIP, we gradually

Fig. 1. Visual comparison of our method with recent state-of-the-art methods. Other methods suffer from contrast degradation and noise artifacts. our method has the best visual perception.

introduce semantic information in the frequency domain space of multiple wavelet transform decompositions, construct a multilevel semantic guidance network to alleviate the difficulty of multi-modal feature alignment, and impose multilevel conditional constraints on the diffusion process to achieve metric-friendly and perceptually oriented enhancement. In addition, we combine the wavelet transform and Fourier transform to construct a high-frequency hybrid space with significant perceptual capabilities. Appearance restoration of degraded images is explored from a spectral perspective, thus further avoiding the generative diversity of diffusion models. Extensive experimental results on public benchmark datasets show that CFWD significantly improves image quality assessment up to state-of-the-art while also providing better visualization.

In summary, the contribution of this paper can be summarised as follows:

- We propose the method of CLIP-Fourier Guided Wavelet Diffusion (CFWD). This is the first successful introduction of multi-modal into the diffusion model-based low-light image enhancement work, which has a more realistic visual perception enhancement performance and a more stable generation effect.
- To further enhance the conditional guidance, we designed a multi-level visual-language guidance network by combining frequency domain space and multi-modal for the first time. It effectively mitigates the multi-modal feature alignment problem caused by image corruption by gradually introducing visual-language information in the frequency domain in combination with the wavelet diffusion process. Meanwhile, the multilevel guidance of the enhancement process is achieved, which significantly improves the metric and visual perception.
- We construct high-frequency hybrid spaces with significant perceptual capabilities by exploring the effective combination of wavelet transform and Fourier transform. Effective constraints on the diversity of diffusion model generation are achieved, and the enhancement performance is effectively improved.

The remainder of this paper is structured as follows. In Section II, the related works are discussed. Section III explains the conventional conditional diffusion model. In Section IV, the proposed novel model method is described in detail. The relevant experimental setup and results are shown in Section V. Section VI is the conclusion.

II. Related Works

A. Traditional Approaches

Low-light image enhancement has received extensive attention from researchers as an important support for various downstream tasks [1], [3], [26]. Traditional low-light image enhancement techniques mainly focus on constructing physical models using two types of methods, adaptive histogram equalisation [4] and Retinex theory [5], which are processed by optimizing the parameter information of the image itself. The former class of algorithms optimizes pixel brightness based on the idea of histogram equalization, while the latter class of methods obtains the desired reflectance map (i.e., the normal image) by estimating the light from the low-light input and removing the effect of the estimated light. For example, [27] achieved enhancement of non-uniform images by balancing detail and naturalness through double logarithmic transformation. [7] proposed a weighted variational model using regularisation terms to estimate the image illumination component and the reflection image. [28] used probing the maximum value in the RGB channel to estimate the illuminance of each pixel and subsequently enhanced the low-light image using a manually designed structural prior.

B. Deep Learning Approaches

The rapid development of deep learning has also triggered the enthusiasm of researchers to explore the field of low-light image enhancement. Numerous low-light enhancement algorithms through data-driven enhancement have been proposed one after another [15], [16], [29]–[31]. Lore et al. [32] proposed LLNet, the first network that applies deep learning to image enhancement, which is trained on degraded images through an encoder-decoder architecture. HDR-Net [33] combines deep networks with the ideas of bilateral grid processing and local affine color transformations with pairwise supervision. [18] proposed Retinex-Net, which first introduced Retinex theory to deep learning and constructed an end-to-end image decomposition algorithm. Zhang et al. [34] proposed the KinD method to improve the problem of producing unnatural enhancement results in Retinex-Net by introducing training loss and adjusting the network architecture. Enlightengan [14] used a generative inverse network as the main framework and was first trained using unpaired images. [12] constructed pixel level by stepwise derivation of the curve estimation convolutional neural network and designed a series of zero-reference training loss functions. [19] utilizes a signal-to-noise ratio aware transformer and a CNN model with spatially varying operations for recovery. Although all these methods have achieved remarkable results, they still face significant challenges in terms of generation quality and enhanced generalization performance due to the lack of effective supervision and efficient reconstruction of the content.

Furthermore, Efficient cross-modal learning has opened up new ideas for computer vision and has been greatly developed. Radford et al. [35] proposed to learn a priori knowledge from large-scale image-text data pairs in order to construct a visual language model CLIP for efficient image classification and task migration with zero-sample training. [36] efficiently

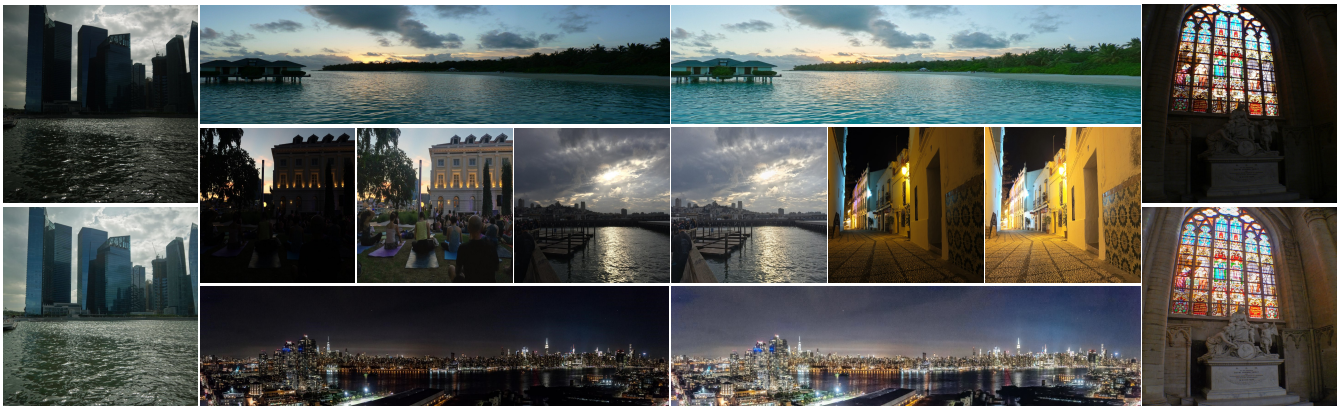

Fig. 2. Representative visual examples by enhancing low-light images using CFWD. All of these images have either 2k resolution or 4k resolution.

performed region enhancement on backlit images by iteratively learning the prompt text from a frozen pre-trained CLIP model. To the best of our knowledge, compared to other methods, we are the first to successfully introduce multi-modal learning in a diffusion model-based low-light image enhancement method and achieve significant performance improvements.

C. Diffusion Models Approaches

Recently, diffusion-based generative models [37] have achieved amazing results with the exploration of many researchers. Meanwhile, low-level visual tasks [38]–[42] have also gained significant progress as a result. Saharia et al. [23] adopt a direct cascading approach, integrating low-resolution measurements and latent codes as inputs to train conditional diffusion models for restoration. WeatherDiff [22] introduces a block-based diffusion model aimed at recuperating images taken in adverse weather conditions, employing guidance across overlapping blocks during the inference stage.

Moreover, for low-light image enhancement, researchers have also recently favoured diffusion model-based approaches. Fei et al. [24] utilize the a priori knowledge embedded in a pre-trained diffusion model to address linear inverse problems effectively. Jiang et al. [25] 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. [43] introduced a diffusion model with a global structure-aware regularisation scheme for the enhancement of degraded images. Different from CFWD, the existing diffusion model approach does not allow for effective guidance and supervision during the enhancement process, leading to unnatural colours and numerous noises during inference. This seriously affects human visual perception and downstream task applications.

III. Preliminary

Diffusion models [37], [44] 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.

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 computing x_0.

The reverse inference process is to recover the original data from Gaussian noise. In contrast to the forward diffusion process, The reverse inference process relies on optimising the noise predictor to iteratively remove the noise and recover the data until the randomly sampled noise $\hat{x}_T \sim N(0, I)$ becomes clean data $\hat{x}_0$. Formulated as:

$$p_\theta(\hat{x}_0, \cdots, \hat{x}_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(\hat{x}_{t-1}|\hat{x}_t), \tag{3}$$

$$p_\theta(\hat{x}_{t-1}|\hat{x}_t) = N(\hat{x}_{t-1}; \mu_\theta(\hat{x}_t, t), \sigma_t^2 I), \tag{4}$$

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:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}(\hat{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\hat{x}_t, t)), \tag{5}$$

where ϵ_θ is a function approximator intended to predict ϵ from $\hat{x}_t$, $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t$=$\prod_{i=1}^{t} \alpha_i$.

IV. Method

As shown in Fig. 3, inspired by [25], our proposed method employs the wavelet diffusion model as a generative framework to reduce the consumption of computational resources.

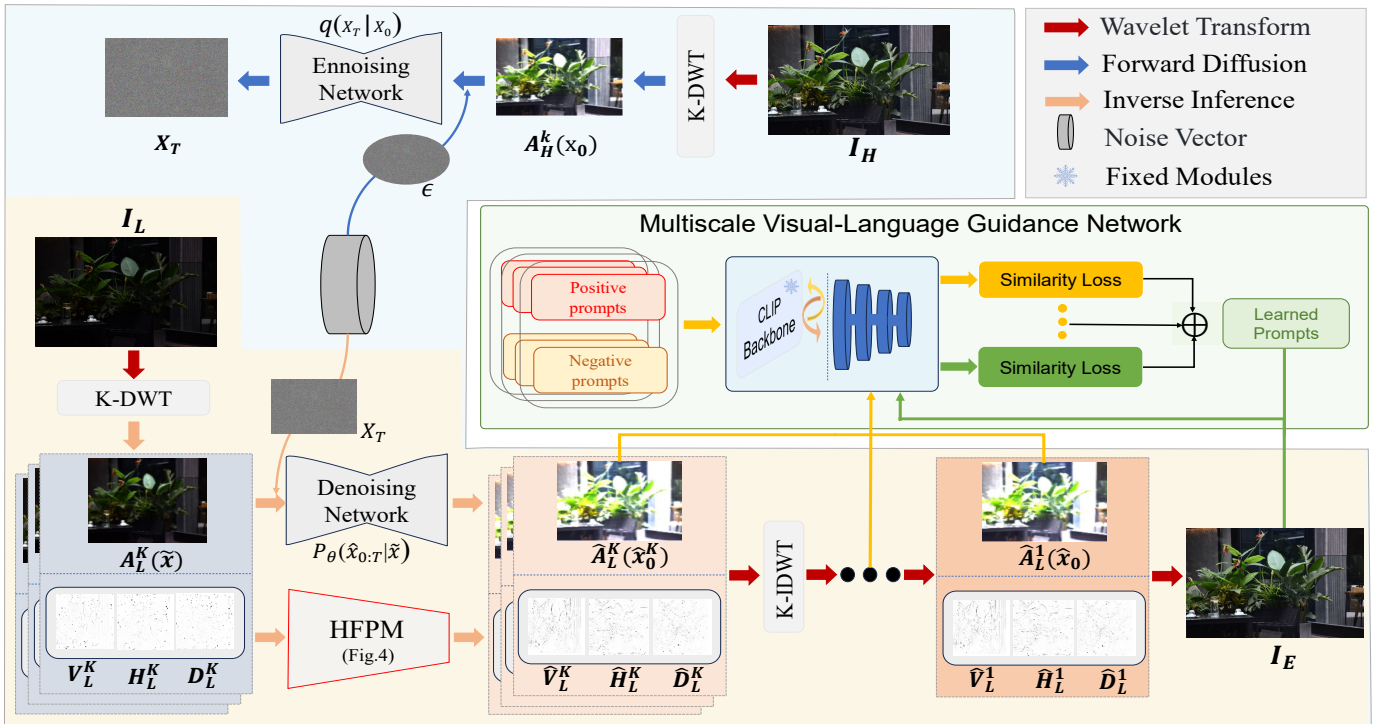

Fig. 3. The overall workflow of our proposed CFWD. It first transforms the low-light input I_L and normal image I_H to the wavelet low-frequency domain (A) for diffusion inference via the K-discrete wavelet transform (K-DWT). We embed a multiscale visual guidance network to iteratively perform appearance guidance and content constraints by combining multiple wavelet domains in the inference process. In addition, the decomposed three high-frequency information $\{V_L, H_L, D_L\}$ we effectively augment by a high-frequency perception module (HFPM). Finally, the final enhancement result I_E is obtained by inverse discrete wavelet transform (K-IDWT).

Meanwhile, we implement iterative guidance of the diffusion process to drive the appearance enhancement by effectively combining the visual-language and wavelet domains at multiple levels, which effectively mitigates the feature alignment difficulties of the visual-language model in the low-light image enhancement task. In addition, we explore the advantageous combination of wavelet transform and Fourier transform to construct a high-frequency perception module to guide the content reconstruction of diffusion models and bridge the gap between degraded and normal domains. Through the effective combination of multi-modal, frequency domain and diffusion models, we achieve high-quality visual enhancement effects and metric results. In this section, we first introduce the generative framework of this paper, i.e., the wavelet diffusion model, and then analyse in detail the multiscale visual-language guidance network and the high-frequency perception module.

A. Wavelet Diffusion Model

Existing diffusion models require high computational resources and are slow in efficiency. Therefore, we reduce the consumption of computational resources by transferring the diffusion process to the wavelet low-frequency domain via discrete wavelet transform. Specifically, in this part, the low-light image $I_L \in R^{H\times W\times C}$ and the normal image $I_H \in R^{H\times W\times C}$ are decomposed using the multiple discrete wavelet transform (K-DWT), where each time it is decomposed into four subbands:

$$\{A^K, V^K, H^K, D^K\} = \text{K-DWT}(I), \tag{6}$$

Where $A^K \in R^{\frac{H}{2^K}\times\frac{W}{2^K}\times C}$ denotes the low-frequency domain of the image after K-DWT. The V^K, H^K, and D^K denote the high-frequency domain of the image in the vertical, horizontal, and diagonal directions, respectively.

Therefore, each discrete wavelet transform performed on an image is equivalent to downscaling its low-frequency domain to one-fourth of the original image. By shifting the diffusion process to take place in the wavelet low-frequency domain, we can significantly reduce the consumption of computational resources due to the substantial reduction in spatial dimensions.

Furthermore, we constrain the content diversity of the sampling process by performing forward diffusion in the wavelet low-frequency domain A_H^K of the normal image I_H and using the wavelet low-frequency domain A_L^K of the degraded image I_L as a conditional guide. Accordingly, Eq. 3 can be rewritten as:

$$p_\theta(\hat{x}_{0:T}|\tilde{x}) = p(\hat{x}_T)\prod_{t=1}^{T} p_\theta(\hat{x}_{t-1}|\hat{x}_t, \tilde{x}). \tag{7}$$

B. Multiscale visual-language Guidance Network

Most of the existing low-light image enhancement algorithms reconstruct the appearance by image-level supervision through a single modality, which leads to difficulties in content reconstruction and significant degradation of the visual quality

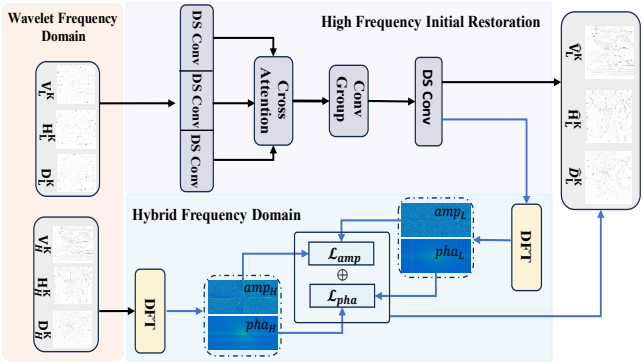

Fig. 4. Detailed architecture of our proposed High Frequency Perception Module (HFPM). DS Conv denotes depth-wise separable convolution, and DFT denotes Discrete Fourier Transform.

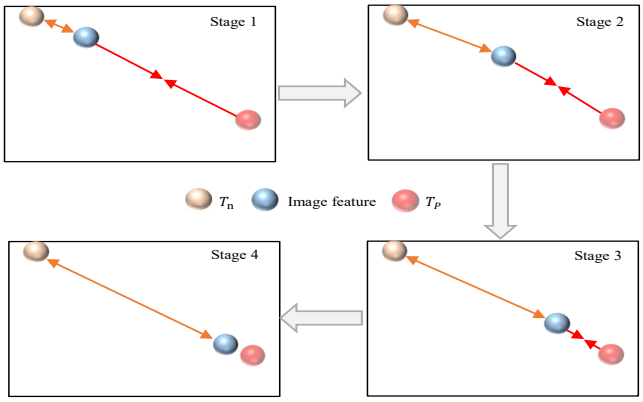

Fig. 5. The multiscale visual-language guidance network gradually promotes the alignment of image features with the positive prompts T_p and continuously moves away from the negative prompts T_n. Stage 1 indicates without visual-language guidance.

of the enhancement process. Meanwhile, simply applying visual-language models in low-level visual tasks does not obtain good performance. This may be due to their inability to capture fine-grained gaps in multi-modal semantics in degraded images, resulting in difficulties in aligning image features with text features.

Therefore, we explore a combined frequency-domain diffusion and multi-modal approach to appearance guidance. The visual-language prompts are used in conjunction with the diffusion model to guide the appearance reconstruction of the wavelet domain of the image. Then, the enhancement results are used for multilevel semantic guidance to promote feature alignment between the image and the visual-language prompts, reaching a two-way iterative optimization effect. The image A_L^K is first combined with visual-language prompts during the diffusion process, then performing coarse-grained feature alignment to obtain preliminary enhancement results $\hat{A}_L^K$. Using $\hat{A}_L^K$ after initially bridging the gap between the weak and normal light domains of I_L, we iteratively instruct its multiple wavelet low-frequency domains $\hat{A}_L^k (k \in [1, K-1])$ with the visual-language positive prompts T_p and negative prompts T_n, expecting the low-light image to be enhanced in the direction of positive prompt T_p and away from negative prompt T_n. As shown in Fig. 5, when we set the wavelet transform scale $K = 2$, through multi-scale semantic iterative guidance, the image is gradually enhanced in the desired direction. This further promotes the feature alignment between the image and the positive prompt T_p and keeps moving away from the negative prompt T_n, realizing bidirectional appearance recovery.

We achieve alignment between images and prompt text features by freezing the latent space of the pre-trained visual-language model CLIP. By driving appearance recovery through visual-language prompts $\{T_p, T_n\}$, we significantly improve the contrast and illumination of the image and achieve stable sampling of the diffusion model. In addition, this section exploits cosine similarity to optimize network training, which can be formulated as follows:

$$\mathcal{L}_{Similarity\_1} = \sum_{k=1}^{K} \Big(\frac{\cos(\Phi_{image}(\hat{A}_L^k), \Phi_{text}(T_n))}{\cos(\Phi_{image}(\hat{A}_L^k), \Phi_{text}(T_p))} + \cos(\Phi_{image}(\hat{A}_L^k), \Phi_{text}(T_p)) \Big), \tag{8}$$

where Φ_{text} is the text encoder, and Φ_{image} is the image encoder. Along with the visual-language guidance, we use inverse discrete wavelet transform to recover the image until the final enhancement result I_E is obtained. At the same time, we employ the learned prompts [36] set to perform fine-grained multi-modal feature alignment on the final enhancement result, further expecting the enhancement result to reduce the distance from the target image, i.e.:

$$\mathcal{L}_{Similarity\_2} = \frac{e^{\cos(\Phi_{image}(I_E), \Phi_{text}(T_n))}}{\sum_{i \in \{T_p, T_n\}} e^{\cos(\Phi_{image}(I_E), \Phi_{text}(T_i))}}. \tag{9}$$

Thus, we can generalize the multiscale visual-language guidance loss as:

$$\mathcal{L}_{vlg} = \mathcal{L}_{Similarity\_1} + \mathcal{L}_{Similarity\_2}. \tag{10}$$

C. High Frequency Perception Module

Diffusion models have strong generative diversity, which becomes a limitation of algorithm performance for image enhancement and restoration tasks. Most of the current low-light image enhancement algorithms based on the diffusion model rely on image-level supervision with content reconstruction losses such as MSE and SSIM to achieve stable sampling of content. However, this does not provide significant content reconstruction of degraded images, which leads to content missing and visual degradation. Therefore, in order to further constrain the diffusion model, it is necessary to avoid generating content diversity and achieve visually oriented enhancement. Inspired by [45], we explore the restoration of image high-frequency information from a frequency domain perspective.

The high-frequency perception module designed in this paper is shown in Fig. 4. Compared with the low-frequency information, the high-frequency information generated by the discrete wavelet transform contains only the details and contours of the image, which can reduce the content interference for the Fourier transform and increase the ability to perceive the details of the image. Thus, we double-transform the image high frequency to construct the hybrid frequency domain space. We first perform detail enhancement [25] on the wavelet high-frequency information generated from the low-light image I_L to obtain more contour structures and image parameters. Specifically, three high-frequency subbands $\{V_L^K, H_L^K, D_L^K\}$ are feature-extracted using depth-wise separable convolutions, and then the detail contours of D are enhanced using V, H combined with cross-attention. Subsequently, the enhanced three high-frequency subbands $\{\hat{V}_L^K, \hat{H}_L^K, \hat{D}_L^K\}$ are obtained by dilation convolutions [46] and depth-wise separable convolutions. After detail enhancement of the high-frequency information of I_L, we perform discrete Fourier transform $\mathrm{DFT}(\cdot)$ on $\{\hat{V}_L^K, \hat{H}_L^K, \hat{D}_L^K\}$ and $\{V_H^K, H_H^K, D_H^K\}$ obtained by the decomposition of the normal image I_H to obtain the spectrum, i.e.:

$$amp_L, pha_L = \mathrm{DFT}(\{\hat{V}_L^K, \hat{H}_L^K, \hat{D}_L^K\}), \tag{11}$$

$$amp_H, pha_H = \mathrm{DFT}(\{V_H^K, H_H^K, D_H^K\}), \tag{12}$$

where amp, pha denote the amplitude and phase of the image, respectively.

To further obtain an enhancement that is consistent with human perception, the method proposed in this paper employs the L_1 loss to minimize the information difference between the high-frequency information spectrograms of normal and low-light images:

$$\mathcal{L}_{spectral} = \vartheta_1 \mathcal{L}_{amp} + \vartheta_2 \mathcal{L}_{pha}, \tag{13}$$

$$\mathcal{L}_{amp} = \frac{1}{K} \sum_{i=1}^{K} \parallel amp_L^i - amp_H^i \parallel_1, \tag{14}$$

$$\mathcal{L}_{pha} = \frac{1}{K} \sum_{i=1}^{K} \parallel pha_L^i - pha_H^i \parallel_1, \tag{15}$$

where ϑ_1 and ϑ_2 are the weighting parameters for the amplitude and phase losses, and i is the scale of the current wavelet transform.

D. Model Training

In CFWD, the loss function can be divided into three main parts: diffusion loss, multi-scale semantic guided loss and content reconstruction loss. Among them, diffusion loss is used to optimize the noise prediction of the diffusion model. In order to initially constrain the content diversity, this paper shifts the diffusion process to the wavelet low-frequency domain to carry out and minimize their L2 distances. Accordingly, the objective function is denoted as:

$$\begin{aligned} \mathcal{L}_{diff} = & E_{t\sim[1,T]} E_{x_0\sim p(x_0)} E_{z_t\sim N(0,I)} \\ & \parallel \epsilon_t - \epsilon_\theta(x_t, \tilde{x}, t) \parallel^2 + ||\hat{A}_L^K - A_H^K||^2. \end{aligned} \tag{16}$$

For content reconstruction loss, in addition to optimizing the spectral loss of details, we perform content reconstruction by combining MSE loss and SSIM loss to minimize the content difference between the recovered image I_L content and the reference image I_H content, i.e. :

$$\begin{aligned} \mathcal{L}_{content} = & \sum_{l=0}^{4} \gamma_l \parallel \Phi_{image}^l(I_E) - \Phi_{image}^l(I_H) \parallel^2 \\ & + (1 - SSIM(I_E, I_H)), \end{aligned} \tag{17}$$

where γ_l is the weight of layer l of the image encoder in the ResNet101 CLIP model.

Accordingly, by combining multiple losses, we significantly enhance the model performance and obtain a satisfactory visual perception, with the total loss denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \mathcal{L}_{vlg} + \mathcal{L}_{spectral} + \mathcal{L}_{content}. \tag{18}$$

V. EXPERIMENTS

A. Experimental Settings

Dataset. Our network is trained and evaluated on the LOLv1 dataset [18], which 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. In addition, we employ two other real-world pairwise datasets, LOLv2-Real_captured [47], and LSRW [48], to evaluate the performance of our proposed network. Specifically, the LOLv2-Real_captured dataset contains 689 low/normal light image pairs for training and 100 for testing. Most low-light images were collected by varying the exposure time and ISO and fixing other camera parameters. The LSRW dataset contains 5,650 image pairs captured in a variety of scenarios. 5,600 image pairs were randomly selected as the training set, and the remaining 50 pairs were used for evaluation. To evaluate the generalization ability of the proposed method in this paper, we tested our method on the BAID [49] test dataset, which consists of 368 backlit images with 2K resolution. In addition, we also tested on two unpaired datasets, LIME [28] and DICM [50].

Implementation Details. We implemented our method with PyTorch on two NVIDIA RTX 3090 GPUs. The network was set up with a total of 2×10^5 iterations, using the Adam optimizer, with the initial learning rate set to 1×10^{-4}, and the batch size and patch size set to 16 and 256×256, respectively.

Evaluation Metrics. For the real-world paired datasets we tested, we used two full-reference distortion measures, PNSR and SSIM [51], as well as two perceptual metrics, LPIPS [52] and FID [53], to evaluate the performance and visual satisfaction of our approach. Higher PSNR or SSIM implies more realistic restoration results, while lower LPIPS or FID indicates higher quality details, brightness and hue. In addition, for the unpaired datasets LIME and DICM, we used three non-reference perceptual metrics: NIQE [54], BRISQUE [55], and PI [56] to evaluate the visual quality of the enhancement results. The lower the metrics, the better the visual quality.

Comparison Methods. To verify the effectiveness of the method proposed in this paper, we compared it with the

TABLE I
QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON LOLV1 [18] , LOLV2-REAL_CAPTURED [47] , AND LSRW DATASETS [48]. THE BEST AND SECOND PERFORMANCE ARE MARKED IN RED AND BLUE, RESPECTIVELY.

Methods	Reference	LOLv1				LOLv2-Real_captured				LSRW			
		PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PNSR↑	SSIM↑	LPIPS↓	FID↓
RetinexNet	BMVC'18	26.316	0.844	0.219	48.037	17.715	0.652	0.436	133.905	15.609	0.414	0.454	108.350
DSLR	TMM'20	14.816	0.572	0.375	104.428	17.000	0.596	0.408	114.306	15.259	0.441	0.464	84.930
DRBN	CVPR'20	16.774	0.462	0.417	126.266	18.466	0.768	0.352	89.085	16.734	0.507	0.457	80.727
Zero-DCE	CVPR'20	14.861	0.559	0.385	87.270	18.194	0.649	0.390	84.123	15.858	0.454	0.421	65.690
MIRNet	ECCV'20	24.138	0.830	0.250	69.179	20.020	0.820	0.233	49.108	16.470	0.477	0.430	93.811
Zero-DCE++	TPAMI'21	14.682	0.472	0.407	87.552	17.461	0.490	0.427	81.727	16.210	0.457	0.431	59.959
EnlightenGAN	TIP'21	17.483	0.651	0.390	95.028	18.676	0.678	0.364	84.044	17.081	0.470	0.420	69.184
ReLLIE	ACM MM'21	11.437	0.482	0.375	95.510	14.400	0.536	0.334	79.838	13.685	0.422	0.404	65.221
RUAS	CVPR'21	16.405	0.499	0.382	102.013	15.351	0.495	0.395	94.162	14.271	0.461	0.501	78.392
DDIM	ICLR'21	16.521	0.776	0.376	84.071	15.280	0.788	0.387	76.387	14.858	0.486	0.495	71.812
CDEF	TMM'22	16.335	0.585	0.407	90.620	19.757	0.630	0.349	74.055	16.758	0.465	0.399	62.780
SCI	CVPR'22	14.784	0.526	0.392	84.907	17.304	0.540	0.345	67.624	15.242	0.419	0.404	56.261
URetinex-Net	CVPR'22	19.842	0.824	0.237	52.383	21.093	0.858	0.208	49.836	18.271	0.518	0.419	66.871
SNRNet	CVPR'22	24.609	0.841	0.262	56.467	21.480	0.849	0.237	54.532	16.499	0.505	0.419	65.807
Uformer	CVPR'22	19.001	0.741	0.354	109.351	18.442	0.759	0.347	98.138	16.591	0.494	0.435	82.299
Restormer	CVPR'22	20.614	0.797	0.288	72.998	24.910	0.851	0.264	58.649	16.303	0.453	0.427	69.219
Palette	SIGGRAPH'22	11.771	0.561	0.498	108.291	14.703	0.692	0.333	83.942	13.570	0.476	0.479	73.841
UHDFour	ICLR'23	23.093	0.821	0.259	56.912	21.785	0.854	0.292	60.837	17.300	0.529	0.443	62.032
CLIP-LIT	ICCV'23	12.394	0.493	0.397	108.739	15.262	0.601	0.398	100.459	13.483	0.405	0.425	77.065
NeRCo	ICCV'23	22.946	0.785	0.311	76.727	25.172	0.785	0.338	84.534	19.456	0.549	0.423	64.555
WeatherDiff	TPAMI'23	17.913	0.811	0.272	73.903	20.009	0.829	0.253	59.670	16.507	0.487	0.431	96.050
GDP	CVPR'23	15.904	0.540	0.431	112.363	14.290	0.493	0.435	102.416	12.887	0.362	0.412	76.908
GSAD	NeurIPS'23	27.629	0.876	0.188	43.659	28.805	0.894	0.201	41.456	19.418	0.542	0.386	57.219
WCDM	TOG'23	26.316	0.844	0.219	48.037	28.875	0.874	0.203	45.395	19.281	0.552	0.350	45.294
CFWD(Ours)	-	29.185	0.872	0.197	40.987	29.855	0.891	0.193	34.814	19.566	0.572	0.374	47.606

TABLE II
QUANTITATIVE COMPARISON OF 2K RESOLUTION BACKLIGHT IMAGES FROM THE BAID [49] DATASET.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓
Zero-DCE++	16.021	0.832	0.240	47.030
EnlightenGAN	17.957	0.866	0.125	47.045
SCI	16.639	0.768	0.197	41.458
SNRNet	17.325	0.754	0.398	63.463
CLIP-LIT	21.611	0.883	0.159	27.926
UHDFour	18.541	0.713	0.319	36.025
WCDM	26.042	0.915	0.130	15.870
CFWD(Ours)	26.918	0.917	0.118	14.852

TABLE III
QUANTITATIVE COMPARISON ON LIME [28] AND DICM [50] DATASETS. OUR METHOD PERFORMS THE BEST CONSISTENTLY.

Metheds	DICM			LIME		
	NIQE↓	BRISQUE↓	PI↓	NIQE↓	BRISQUE↓	PI↓
DRBN	4.369	30.708	3.782	4.562	29.564	3.573
Zero-DCE	3.414	36.452	2.911	3.771	18.481	2.759
MIRNet	4.021	22.104	3.391	4.378	28.623	2.998
RUAS	5.119	41.897	4.127	4.702	29.601	3.479
DDIM	3.899	19.787	3.013	3.899	24.474	3.059
EnlightenGAN	3.439	14.175	2.719	3.656	14.854	2.832
SCI	3.519	25.289	2.824	4.163	17.094	2.908
URetinex-Net	4.774	24.544	3.565	4.694	29.022	3.313
SNRNet	4.070	26.179	3.926	5.691	34.187	4.636
CLIP-Lit	3.557	26.991	2.589	3.989	19.422	2.813
NeRCo	3.329	19.586	2.890	3.803	21.164	2.888
UHDFour	4.231	13.174	3.186	4.627	15.930	3.344
GDP	4.358	19.294	2.887	4.186	22.022	3.109
GSAD	3.735	20.296	2.894	4.578	26.356	3.492
WCDM	3.364	15.862	2.364	3.597	14.474	2.605
CFWD(Ours)	3.322	10.955	2.699	3.568	10.141	2.686

State-of-the-art methods in recent years, including RetinexNet [18], DSLR [57], DRBN [58], Zero-DCE [12], Zero-DCE++ [59], MIRNet [60], EnlightenGAN [14], ReLLIE [61], RUAS [62], DDIM [44], SCI [63], URetinex-Net [64], SNRNet [19], Palette [65], Uformer [66], Restormer [67], CDEF [68], UHDFour [69], CLIP-LIT [36], NeRCo [70], WeatherDiff [22], GDP [24], WCDM [25] and GSAD [43].

B. Results

Quantitative Comparison. Firstly, we compare our method with all state-of-the-art methods on the LOLv1 [18], LOLv2-Real_captured [47], and LSRW [48] test sets. As shown in Table I, our method achieves state-of-the-art quantization performance in several metrics compared to all methods. In particular, the significant improvements in PSNR and FID provide compelling evidence for the superior perceived quality of our method. Specifically, for two distortion metrics, our method obtains all firsts in PSNR evaluation, achieving performance improvements of 1.556dB, 0.98dB, and 0.11dB in

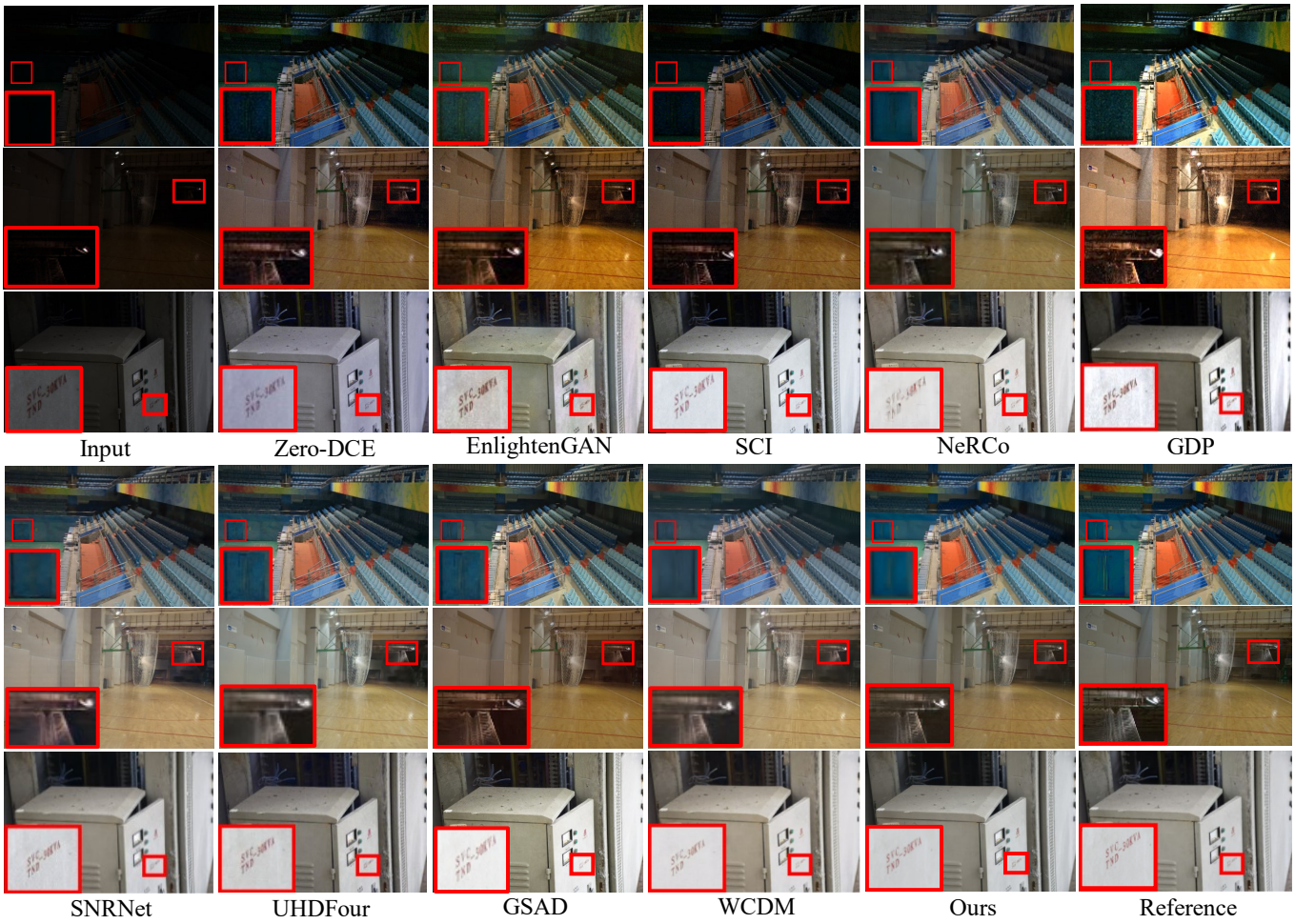

Fig. 6. Visual comparison of our method with State-of-the-art methods on LOLv1 [18](row 1), LOLv2-Real_captured [47](row 2), and LSRW [48](row 3) datasets from various years in recent years. Our method is closer to a normal image, best viewed by zooming in.

the LOLv1, LOLv2-Real_captured, and LSRW datasets, respectively. Furthermore, our method achieves the second-best SSIM quantisation performance on the LOLv1 and LOLv2-Real_captured datasets. Compared to the third-place WCDM, our method has a significant improvement of 0.028 (LOLv1) and 0.017 (LOLv2-Real_captured), respectively, while for the first-place GSAD, we only have a small difference of 0.004 and 0.003. For two perceptual metrics (i.e., LPIPS and FID), our method meets the quantitative criteria on the LOLv2-Real_captured dataset and is well ahead of competing methods. We are also significantly competitive on the LOLv1 and LSRW datasets, obtaining three second-place as well as one first-place quantitative performances. This indicates that the method proposed in this paper can generate recovered images with satisfactory visual quality, further demonstrating the effectiveness of our method. Table II also provides a quantitative comparison of some state-of-the-art methods on the BAID [49] test dataset. From the evaluation metrics, our method outperforms all the state-of-the-art methods, which indicates that our proposed method is more effective in terms of generalisation ability and high-resolution low-light image restoration.

Meanwhile, we performed evaluation comparisons with competing methods on two unpaired datasets LIME [28] and DICM [50] to validate the effectiveness and generalization of our method. We evaluated the effectiveness of our method in terms of visual quality by combining three non-reference perceptual metrics, NIQE, BRISQUE and PI, with lower metrics resulting in better visual quality. As shown in Table III, our method meets the quantification criteria on both datasets compared to other competing methods. Specifically, we obtain the best performance for all quantitative assessments for both NIQE and BRISQUE, while for the PI metrics, we also have the second-best results. This further demonstrates the better generalisability of our approach in real-world scenarios and enhancements that are more in line with human visual perception.

Visual Comparison. Fig. 6 is shown to compare our method with State-of-the-art methods on the paired dataset. The images in rows 1-3 are selected from the LOLv1, LOLv2-Real_captured, and LSRW test sets. The visualization of the BAID dataset is then shown in Fig. 7. Through these comparisons, it is easy to see that previous methods seem to suffer from incorrect exposure, color distortion, noise ampli-

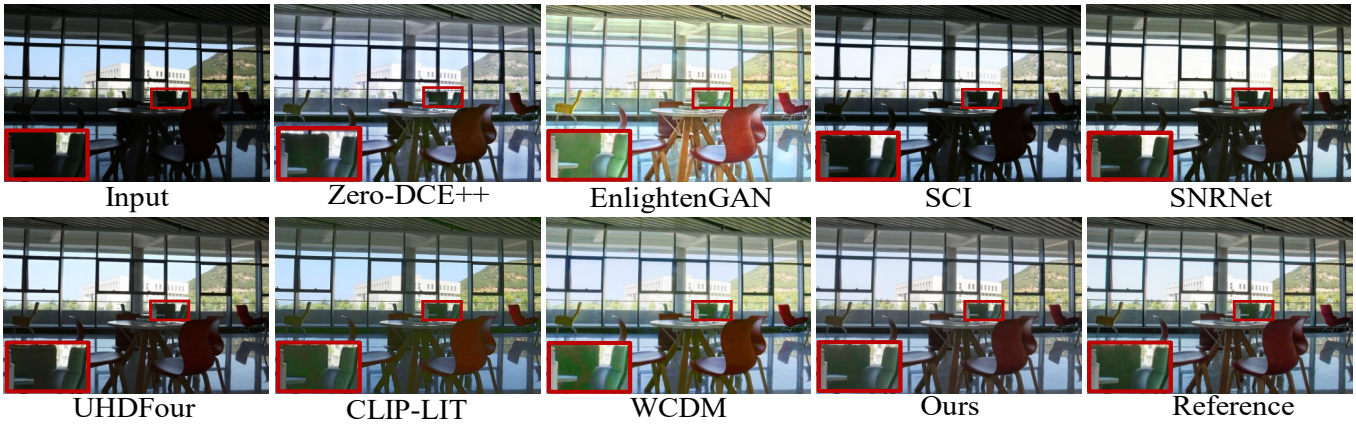

Fig. 7. Visual comparison of 2K resolution backlight images of our method and competing methods on BAID [49] test set. It is best viewed by zooming in.

fication, or artefacts, which affect the overall visual quality. For example, EnlightenGAN and GDP suffer from generation artefacts and noise amplification, while SNRNet and WCDM suffer from color distortion. In addition GSAD fails to produce similar colours and contrast as the reference image. In contrast, our method consistently produces visually pleasing results with improved color and brightness without overexposure or underexposure. We attribute this to the improved appearance of the multilevel visual-language guidance network. At the same time, CFWD effectively improves contrast, reconstructs sharper details, and brings the visual effect closer to the original image due to the effective constraints imposed by the high-frequency perceptual module on the content structure.

The visual presentation of the DICM and LIME datasets is shown in Fig. 8. It is clear that our model skilfully adjusts the illumination conditions to optimally improve the contrast of the degraded images while vigilantly avoiding overexposure. This successful balance confirms the generalisability of our proposed method to unseen scenes as well as the satisfaction with the visual results.

C. Ablation Study

To verify the validity of the proposed method, in this subsection, we will conduct an ablation study of the multiscale visual-language guidance network and the high-frequency perception module, and explore the optimal parameter pairing of the network. All the ablation studies are performed entirely on the LOLv1 dataset.

Multiscale visual-language Guidance Network. Benefitting from the efficient visual-language prior to CLIP, our method can learn different modalities and thus produce better perceptual and metric results. In order to investigate the effect of the level M of the visual-language guidance network on our method, we fixed the number of wavelet transforms to 2 and verified its effectiveness by gradually increasing the level of visual-language guiding. As shown in Table IV, when M=0, it indicates that we give up the multimodal learning, and by comparison, we find that after multimodal visual-language guiding, we effectively improve the performance of the network. Meanwhile, with the gradual increase of M, the performance of the network steadily improves. This indicates that multilevel visual-language guidance can iteratively guide the fine-grained alignment of image features with text features during the enhancement process and bring significant network performance improvement.

TABLE IV
RESULTS OF AN ABLATION STUDY AT THE PROMPT NETWORK SCALE.

Prompts Scale	PSNR↑	SSIM↑	LPIPS↓	FID↓
M=0	26.705	0.856	0.227	49.926
M=1	27.809	0.866	0.225	48.501
M=2	28.512	0.871	0.216	43.167
M=3	**29.212**	**0.872**	**0.197**	**40.987**

Hybrid Frequency Domain Perception Module. Due to the obvious differences in the feature information contained in the frequency domain space at different stages, we tested a series of combined experiments on the high-frequency perception module, resulting in three HFPM versions. Specifically, HFPM_v1 uses the wavelet low-frequency domain for Fourier transform to capture image features, HFPM_v2 uses only the high-frequency space of the first wavelet transform to construct a mixed-frequency domain to capture image information, and HFPM_v3 performs Fourier transforms on all the wavelet high-frequency domains obtained from multiple wavelet transforms to form a multi-group mixed frequency domain space. By combining multiple sets of mixed-frequency domain spaces, it can effectively acquire high-frequency features. As shown in the Table V, the performance of the network using the HFPM_v1 version is the worst, which may be due to the fact that the wavelet low-frequency domain contains more structural information, which causes more content loss and feature interference when performing the Fourier transform, resulting in a more chaotic feature learned by the model. In addition, compared with HFPM_v2, HFPM_v3 has better quantitative results, for the wavelet high-frequency domain, we only need the contour and detail information of the image, therefore, with the combination of multi-group mixing space constraints, we can obtain more detail information to constrain the diversity of the diffusion model content.

Fig. 8. Visual comparison on the DICM [50] (row 1), LIME [28] (row 2) datasets among State-of-the-art low-light image enhancement approaches.

TABLE V
ABLATION STUDIES OF THE OPTIMAL EFFECTIVENESS OF OUR HYBRID FREQUENCY DOMAIN PERCEPTION MODULE.

Versions	PSNR↑	SSIM↑	LPIPS↓	FID↓
HFPM_v1	27.638	0.862	0.215	43.193
HFPM_v2	28.282	0.868	0.209	41.185
HFPM_v3	**29.212**	**0.872**	**0.197**	**40.987**

D. Discussion

Despite the excellent performance and visual perception of our proposed low-light image enhancement method, the method still has some non-negligible limitations and goals that need to be further explored. Firstly, the wavelet diffusion model-based low-light enhancement method still has a large computational overhead, which is not conducive to realistic deployment. Second, multiscale visual-language guidance increases the complexity of prompt text design and also carries the risk of augmenting redundant content to some extent. Finally, the loss function required for the enhancement process is more complex, making it difficult to seek the optimal set of weighting parameters.

In the future, we will investigate a more effective diffusion framework based on the above issues and formulate a more model-compliant visual-language learning network to formulate the appropriate visual-language prompts and remove the risk of redundant content. In addition, the further compact design of the loss function will be the core of our exploration, and through the corresponding research, we believe that the proposed method has further performance space.

VI. CONCLUSIONS

We first successfully introduce multimodal into a diffusion model-based approach for low-light image enhancement and propose a wavelet diffusion model based on CLIP and Fourier transform guidance. By combining the generative power of the diffusion model and the visual-language prior to driving the degraded images for appearance restoration, the visual perception and metric performance are significantly enhanced. In addition, we design a novel high-frequency perception module that effectively constrains the diversity of diffusion model-generated content by exploring the advantages of combining the wavelet and Fourier transforms for double transformation, constructing a hybrid frequency-domain space that is acutely aware of the image structure and provides guidance similar to the target result. Extensive experiments conducted on publicly available benchmark datasets show that our method has better stability and generalisability to provide enhancement of degraded images that approximate the reference image.

REFERENCES

[1] K. Dong, Y. Guo, R. Yang, Y. Cheng, J. Suo, and Q. Dai, "Retrieving object motions from coded shutter snapshot in dark environment," *IEEE Transactions on Image Processing*, 2023.

[2] G. Li, Y. Yang, X. Qu, D. Cao, and K. Li, "A deep learning based image enhancement approach for autonomous driving at night," *Knowledge-Based Systems*, vol. 213, p. 106617, 2021.

[3] M. Xue, P. Shivakumara, C. Zhang, Y. Xiao, T. Lu, U. Pal, D. Lopresti, and Z. Yang, "Arbitrarily-oriented text detection in low light natural scene images," *IEEE Transactions on Multimedia*, vol. 23, pp. 2706–2720, 2020.

[4] E. D. Pisano, S. Zong, B. M. Hemminger, M. DeLuca, R. E. Johnston, K. Muller, M. P. Braeuning, and S. M. Pizer, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of Digital imaging*, vol. 11, pp. 193–200, 1998.

[5] E. H. Land and J. J. McCann, "Lightness and retinex theory," *Josa*, vol. 61, no. 1, pp. 1–11, 1971.

[6] J. Park, A. G. Vien, J.-H. Kim, and C. Lee, "Histogram-based transformation function estimation for low-light image enhancement," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1–5.

[7] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2782–2790.

[8] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2828–2841, 2018.

[9] D. Sugimura, T. Mikami, H. Yamashita, and T. Hamamoto, "Enhancing color images of extremely low light scenes based on rgb/nir images acquisition with different exposure times," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3586–3597, 2015.

[10] S. Zhang, N. Meng, and E. Y. Lam, "Lrt: an efficient low-light restoration transformer for dark light field images," *IEEE Transactions on Image Processing*, 2023.

[11] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 769–777.

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

[13] W. Ren, S. Liu, L. Ma, Q. Xu, X. Xu, X. Cao, J. Du, and M.-H. Yang, "Low-light image enhancement via a deep hybrid network," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4364–4375, 2019.

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

[15] K. Zhang, C. Yuan, J. Li, X. Gao, and M. Li, "Multi-branch and progressive network for low-light image enhancement," *IEEE Transactions on Image Processing*, 2023.

[16] Y.-F. Wang, H.-M. Liu, and Z.-W. Fu, "Low-light image enhancement via the absorption light scattering model," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5679–5690, 2019.

[17] Y. Lu and S.-W. Jung, "Progressive joint low-light enhancement and noise removal for raw images," *IEEE Transactions on Image Processing*, vol. 31, pp. 2390–2404, 2022.

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

[19] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 714–17 724.

[20] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[22] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[23] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[24] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9935–9946.

[25] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–14, 2023.

[26] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "Rgb-t salient object detection via fusing multi-level cnn features," *IEEE Transactions on Image Processing*, vol. 29, pp. 3321–3335, 2019.

[27] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE transactions on image processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[28] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on image processing*, vol. 26, no. 2, pp. 982–993, 2016.

[29] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," *Advances in neural information processing systems*, vol. 29, 2016.

[30] L.-W. Wang, Z.-S. Liu, W.-C. Siu, and D. P. Lun, "Lightening network for low-light image enhancement," *IEEE Transactions on Image Processing*, vol. 29, pp. 7984–7996, 2020.

[31] M. Lamba, K. K. Rachavarapu, and K. Mitra, "Harnessing multi-view perspective of light fields for low-light imaging," *IEEE Transactions on Image Processing*, vol. 30, pp. 1501–1513, 2020.

[32] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[33] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

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

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[36] Z. Liang, C. Li, S. Zhou, R. Feng, and C. C. Loy, "Iterative prompt learning for unsupervised backlit image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8094–8103.

[37] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[38] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Image restoration with mean-reverting stochastic differential equations," *arXiv preprint arXiv:2301.11699*, 2023.

[39] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.

[40] M. Ren, M. Delbracio, H. Talebi, G. Gerig, and P. Milanfar, "Image deblurring with domain generalizable diffusion models," *arXiv preprint arXiv:2212.01789*, 2022.

[41] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma, "Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Transactions on Image Processing*, 2023.

[42] L. Liao, W. Chen, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, "Unsupervised foggy scene understanding via self spatial-temporal label diffusion," *IEEE Transactions on Image Processing*, vol. 31, pp. 3525–3540, 2022.

[43] J. Hou, Z. Zhu, J. Hou, H. Liu, H. Zeng, and H. Yuan, "Global structure-aware diffusion process for low-light image enhancement," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[44] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[45] Z. Wang, Z. Yan, and J. Yang, "Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution," *arXiv preprint arXiv:2312.05799*, 2023.

[46] J. Hai, R. Yang, Y. Yu, and S. Han, "Combining spatial and frequency information for image deblurring," *IEEE Signal Processing Letters*, vol. 29, pp. 1679–1683, 2022.

[47] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Transactions on Image Processing*, vol. 30, pp. 2072–2086, 2021.

[48] J. Hai, Z. Xuan, R. Yang, Y. Hao, F. Zou, F. Lin, and S. Han, "R2rnet: Low-light image enhancement via real-low to real-normal network," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103712, 2023.

[49] X. Lv, S. Zhang, Q. Liu, H. Xie, B. Zhong, and H. Zhou, "Backlitnet: A dataset and network for backlit image enhancement," *Computer Vision and Image Understanding*, vol. 218, p. 103403, 2022.

[50] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2d histograms," *IEEE transactions on image processing*, vol. 22, no. 12, pp. 5372–5384, 2013.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[54] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[55] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[56] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[57] S. Lim and W. Kim, "Dslr: Deep stacked laplacian restorer for low-light image enhancement," *IEEE Transactions on Multimedia*, vol. 23, pp. 4272–4284, 2020.

[58] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3063–3072.

[59] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4225–4238, 2021.

[60] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 492–511.

[61] R. Zhang, L. Guo, S. Huang, and B. Wen, "Rellie: Deep reinforcement learning for customized low-light image enhancement," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 2429–2437.

[62] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. 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*, 2021, pp. 10 561–10 570.

[63] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5637–5646.

[64] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. 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*, 2022, pp. 5901–5910.

[65] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[66] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 683–17 693.

[67] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[68] X. Lei, Z. Fei, W. Zhou, H. Zhou, and M. Fei, "Low-light image enhancement using the cell vibration model," *IEEE Transactions on Multimedia*, 2022.

[69] C. Li, C.-L. Guo, M. Zhou, Z. Liang, S. Zhou, R. Feng, and C. C. Loy, "Embedding fourier for ultra-high-definition low-light image enhancement," *arXiv preprint arXiv:2302.11831*, 2023.

[70] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 918–12 927.