This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2026.3684026

Recovering the Unseen: Multi-level Priors-Guided Diffusion for Recovering Brightness-Missing Regions in Extreme Exposure

Xinyue Zhao, Andong Zhang, Zhengyan Xu, Yachao Li, Zaixing He, *Senior Member, IEEE*, Dong Liang, *Senior Member, IEEE*

***Abstract*—In digital photography, images captured under extremely dim or excessively bright lighting often suffer from abnormal exposure, which can lead to irreversible loss of local visual information. While image enhancement has been widely explored, existing approaches rarely succeed in handling extreme exposure scenarios. Unlike moderate exposure issues that merely compress the dynamic range, extreme exposure causes truncation of pixel values, resulting in regions devoid of brightness information—referred to as brightness-missing regions. The absence of reliable cues in these areas poses a major obstacle to both exposure adjustment and detail reconstruction. To address this, we propose a Multi-level Priors-guided Diffusion (MPD) model tailored for extreme exposure restoration. MPD is a unified framework capable of processing both extreme underexposure and overexposure. It leverages three complementary priors—low-level brightness, high-level structural semantics and exposure semantics—within a two-stage pipeline. Initially, pixel-wise brightness correction is performed using low-level brightness-guided dynamic convolution. Subsequently, a diffusion model, guided by structural and exposure semantics, regenerates content in brightness-missing zones to ensure semantic plausibility and natural illumination. Comprehensive experiments confirm MPD's effectiveness and superiority across no-reference and full-reference image quality metrics, as well as in human perceptual evaluations. Source code is available at: https://github.com/zzz-cv/MDP.**

***Index Terms*—Extreme exposure restoration, Semantic consistency, Diffusion models**

I. Introduction

IN digital photography, capturing images under insufficient or excessive illumination frequently results in abnormal exposure. This not only degrades the visual appeal of photographs but also complicates subsequent post-processing tasks. Although significant progress has been made in image enhancement [1]–[3], exposure correction [4]–[6] and high dynamic range (HDR) reconstruction [7]–[11], few methods are capable of effectively handling scenes with extreme exposure. The fundamental distinction lies in the nature of the degradation: non-extreme exposure generally compresses the image's dynamic range, while extreme exposure induces a grayscale truncation phenomenon that permanently removes details and texture, leading to what we term brightness-missing regions, as illustrated by the histograms in Fig. 1.

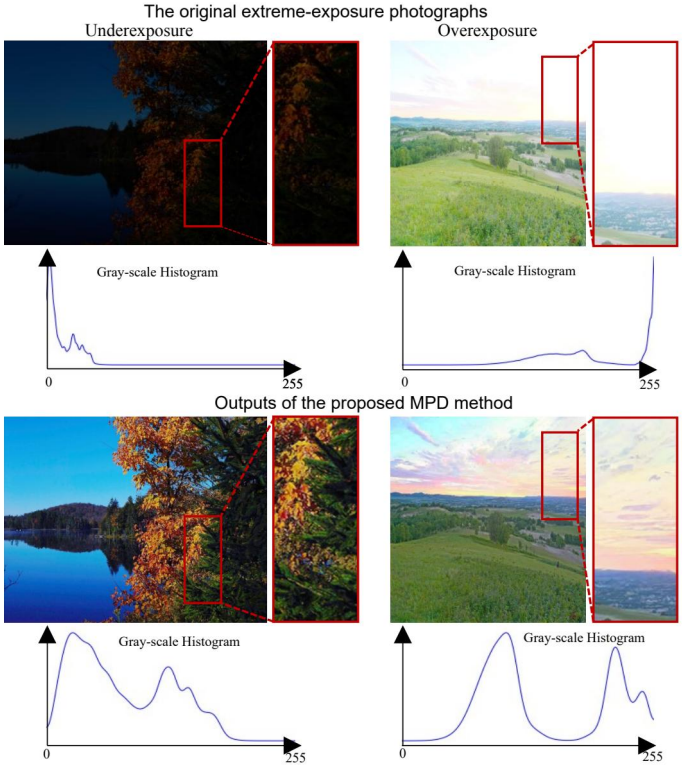


Fig. 1. Original images and Outputs of MPD in extreme underexposure and overexposure digital photography with their corresponding gray-scale histograms. In the histograms of the original images, we can observe that an extreme exposure image not only compresses the images' dynamic ranges but also involves truncating image information, resulting in brightness-missing regions in the images.

The critical challenge in extreme exposure restoration stems from the irreversible loss of photometric information in these brightness-missing regions, where sensor measurements hit the physical limits of the camera's dynamic range. Conventional enhancement techniques [1]–[5], [9]–[11], which typically rely on parametric curve adjustments or local histogram matching, are fundamentally limited as they depend on existing pixel correlations—precisely the information that is destroyed in extreme exposure scenarios. Recent advancements in diffusion models [12]–[14] offer a promising new direction. Unlike conventional methods, diffusion models possess rich prior knowledge learned from large-scale datasets and exhibit powerful generative capabilities through iterative denoising. These

Manuscript received xxxx 2025. This work was supported in part by the National Natural Science Foundation of China under Grants 62272229, 52275547 and 52275514. Xinyue Zhao, Andong Zhang, and Zhengyan Xu contributed equally to this work. (Corresponding author: Zaixing He, Dong Liang.)

Xinyue Zhao and Zaixing He are with the School of Mechanical Engineering, The State Key Lab of Fluid Power and Mechatronic Systems, Zhejiang University

Andong Zhang, Yachao Li, Dong Liang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence.

Zhengyan Xu is with the School of Computer Science and Technology, Beijing Institute of Technology.

characteristics enable them to progressively reconstruct plausible texture details in brightness-missing regions. By framing image restoration as a conditional generation process guided by structural and exposure semantics, diffusion models can infer coherent content in information-truncated areas through multi-step refinement, overcoming the reliance on existing pixel correlations inherent in traditional approaches.

Inspired by this, we propose a Multi-level Priors-guided Diffusion (MPD) model to build a two-stage (first brightness correction and then brightness-missing regions recovery) unified framework for dealing with both extreme underexposure and overexposure. It integrates three types of prior knowledge: low-level brightness, high-level structural semantics and exposure semantics, providing comprehensive guidance for image restoration under extreme exposure conditions.

- *Low-level Brightness Priors* guide an adaptive, pixel-wise brightness correction.
- *Structural Semantics Priors* ensure the recovered content maintains semantic consistency with the original image structure, mitigating empirical bias and illusions in the generative process.
- *Exposure Semantics Priors* harmonize the overall illumination, ensuring the recovered regions blend naturally with the corrected areas to achieve a balanced, natural-looking exposure.

This synergistic guidance mechanism addresses the joint optimization challenge of semantic accuracy and exposure naturalness.

The novelty of the proposed approach lies not merely in the assembly of the priors but in the principled design of a synergistic guidance framework tailored for the distinctive challenge of extreme exposure restoration. First, we define and target the recovery of brightness-missing regions, a problem characterized by irreversible information loss, which necessitates going beyond conventional dynamic range adjustment. Second, we introduce a novel triad of complementary priors that operate in concert: low-level brightness priors enabling pixel-adaptive correction via brightness-aware dynamic convolution; high-level structural semantics priors that enforce contextual consistency through dynamic cross-scale attention between missing regions and known semantics; and crucially; high-level exposure semantics priors that leverage CLIP-based textual prompts to regulate illumination naturalness via cross-modal alignment. This last component represents a distinct innovation for explicitly controlling physical imaging attributes (exposure) within a generative process. Finally, the synergistic interaction between the structural and exposure semantics priors provides a novel solution to the joint optimization problem of semantic accuracy and exposure naturalness in information-truncated areas, a core challenge previously unaddressed in a unified manner.

MPD demonstrates superior performance over current state-of-the-art approaches via extensive experimental comparisons involving visual quality assessments, no-referenced/full-referenced image quality assessments, and human subjective surveys.

II. Related Work

A. Exposure Correction

Exposure correction methods aim to recover images degraded by under- or overexposure, balancing luminance restoration and detail preservation. Recent representative approaches include MSEC [6], which adopts a multi-scale pyramid strategy for coarse-to-fine correction, and LCDPNet [5], which exploits local color distribution cues to distinguish and rectify improperly exposed regions. Beyond these, modern Retinex-based designs [15], [16] further decouple illumination and reflectance to enhance stability under challenging lighting, while sequence modeling and transformer-style architectures improve long-range dependency modeling for more consistent structure recovery. In addition, approaches [17]–[20] prioritize perceptual fidelity via constraint-aware objectives, combining learned priors with explicit color, tone, and gradient regularizers. In the field of HDR reconstruction [7]–[11], single-frame methods such as HDRnet [21] and HDRCNN [22] attempt to correct the brightness of exposure photos, but they often fail to effectively address brightness-missing regions.

The above exposure correction methods adjust images' dynamic range to correct brightness. However, these methods cannot handle extreme exposure conditions and fail to restore brightness in brightness-missing regions under extreme underexposure/overexposure conditions.

B. Diffusion Models in Image Restoration

Diffusion models [12]–[14] are used to generate high-quality images by gradual denoising from random noise, leveraging their priors for various image restoration tasks such as super-resolution [23]–[25], and inpainting [26]–[28]. Diffusion-based image inpainting aims to fill masked image regions using prior knowledge from diffusion processes, achieving restorations that blend seamlessly with the original image context. Uni-paint [27] presents a unified multimodal inpainting framework built on a pretrained diffusion model, enabling text-, stroke-, and exemplar-guided inpainting via few-shot finetuning on the given image. RePaint [26] employs a pre-trained diffusion model to produce high-quality, varied results across facial and generic images without modifying the model's architecture. Despite their successes, the output of these methods can sometimes appear disordered or random.

To reduce the randomness of diffusion-based completion, several recent methods [29]–[31] introduce more explicit structural and semantic constraints into the generation process. Instead of relying on one-shot generation, they usually adopt staged or dual-branch designs that first perform degraded-region localization, semantic drafting, or coarse recovery, and then refine the missing content under structural, semantic, or consistency constraints. Although mainly developed for completion tasks, these methods share somewhat a similar motivation with ours, namely, improving restoration controllability and coherence by coupling generative recovery with explicit priors.

ControlNet [32] enhances generation controllability by adapting the encoder's weights within the diffusion model to learn additional conditions. However, its design objective

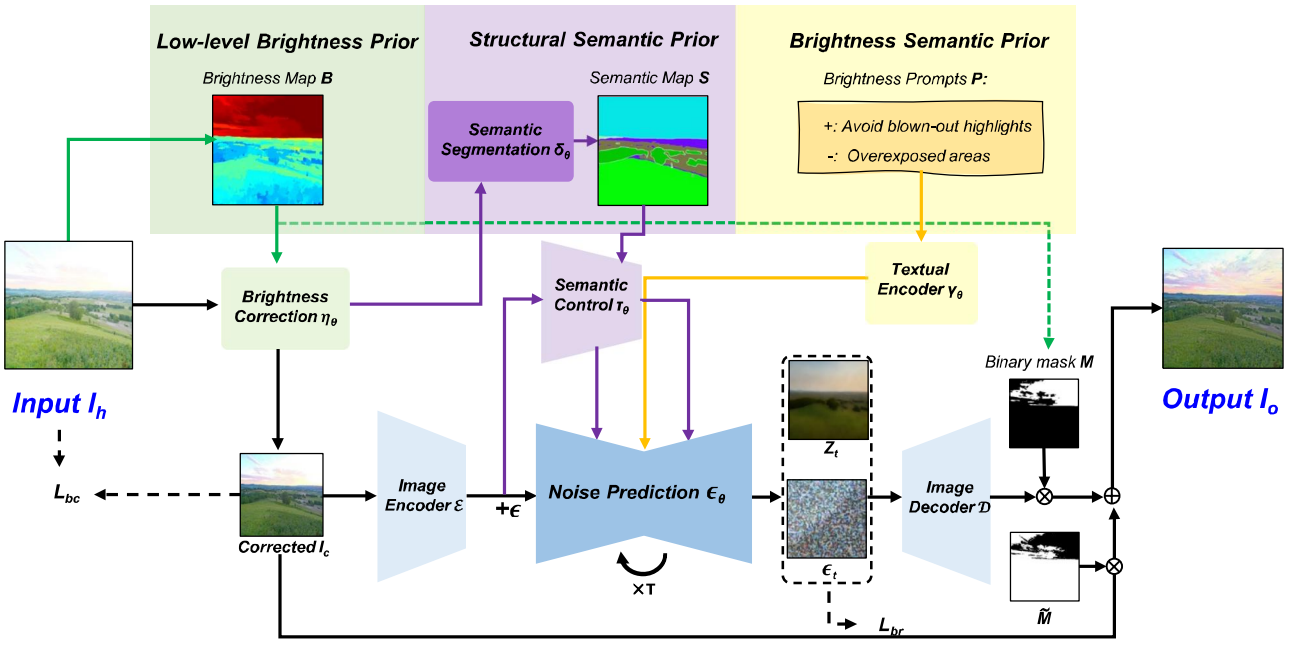


Fig. 2. Overall architecture of MPD. Stage I performs pixel-wise brightness correction through Brightness-aware Dynamic Convolution guided by low-level brightness prior, generating preliminary corrected image I_c. Stage II reconstructs brightness-missing regions using a diffusion model guided by structural semantics prior and exposure semantics prior.

is for conditional guidance in general image generation, and it does not address the problem of brightness-missing region recovery under extreme exposure conditions. The Structural Semantics Priors and Exposure Semantics Priors proposed in this paper are not a simple combination of conditional controls, but are customized for the characteristics of the extreme exposure recovery task: (1) Structural Semantics Priors ensure the structural consistency between the recovered content and the original image through the interaction of semantic segmentation maps and the context of brightness-missing regions, avoiding semantic inconsistency caused by information loss; (2) Exposure Semantics Priors first combine CLIP text prompts (such as "Avoid blown-out highlights") with brightness alignment to directly constrain the exposure naturalness of the generated regions. The synergistic effect of the two addresses the brightness-semantic joint optimization issue not covered by ControlNet.

Under extreme exposure conditions, images suffer from brightness-missing regions, a challenge that traditional brightness correction methods are powerless against. Our method attempts to leverage the reliable prior knowledge and the diffusion model's capability to progressively restore the brightness-missing regions and avoid empirical bias and illusions in the diffusion model, thereby restoring their details and texture coherence with the original image.

III. Proposed Method

A. Preliminary

Diffusion models [12], [33] constitute a class of generative models that learn to approximate the true data distribution $p(z)$ by modeling a transformed distribution $q(z)$ through a Markov chain [34]. The fundamental framework establishes a critical joint distribution:

$$p_\theta(z_{0:T}) = p_\theta^{(T)}(z_T) \prod_{t=0}^{T-1} p_\theta^{(t)}(z_t|z_{t+1}) \tag{1}$$

The forward fixed variational inference distribution is:

$$q(z_{1:T}|z_0) = q^T(z_t|z_0) \prod_{t=1}^{T-1} q^{(t)}(z_t|z_{t-1}, z_0) \tag{2}$$

Then Gaussian parameterization is adopted for both p_θ and q, with one common parameterization form being:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}), \ \forall t \in [1, T] \tag{3}$$

Through recursive application of Eq. 3, we derive the closed-form expression for z_t given z_0:

$$q(z_t|z_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} z_0, (1-\bar{\alpha}_t)\mathbf{I}), \quad \forall t \in [1, T] \tag{4}$$

where the coefficients $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\alpha_t = 1 - \beta_t$. Using reparameterization, we can express Eq. 4 as follows:

$$z_t(z_0, \epsilon) = a_t z_0 + \sigma_t \epsilon \tag{5}$$

where a_t and σ_t are predefined scale factors. This formulation enables direct estimation of the clean image $z_{0|t}$ through a noise prediction network ϵ_θ:

$$z_{0|t} = \frac{1}{a_t}(z_t - \sigma_t \epsilon_\theta) \tag{6}$$

According to Eq. 6, the relationship between z_{t-1} and $z_{0|t}$ can be expressed as follows:

$$z_{t-1} = a_{t-1} z_{0|t} + \sigma_{t-1} \epsilon \tag{7}$$

Based on the above formula, after the T steps of the denoising process, we can reconstruct the clean sample z_0 from the sample z_t. However, the entire denoising process needs to be performed step by step, which leads to a long inference time. Several techniques [35], [36] have been proposed to accelerate this process, such as DDIM [37], which introduces the following scheduling strategy:

$$z_{t-1} = a_{t-1} z_{0|t} + \sigma_{t-1}(\eta_t \epsilon + \sqrt{1-\eta_t^2}\epsilon_\theta) \tag{8}$$

where η_t is an interpolation factor that controls the ratio of the newly introduced noise ϵ.

The reverse process of diffusion models, which involves gradually denoising to reconstruct the original image, offers a promising approach to addressing the challenges of extreme exposure restoration. By iteratively estimating and removing noise, diffusion models can potentially recover missing details in brightness-missing regions caused by extreme exposure.

B. Formulation and Main Idea

Unlike traditional exposure correction or image enhancement tasks, our proposed method focuses on image restoration under extreme exposure scenarios where brightness-missing regions occur. For input images I_h with abnormal brightness, our method first finds a mapping function F to correct the brightness globally, then identifies brightness-missing regions M and utilizes the mapping function G to restore these regions. As shown in Fig. 2, here, we introduce three types of priors: one related to low-level brightness, represented as the brightness map B; another related to image structural semantics, described as the priors S; and the third type is exposure semantics, denoted as exposure prompts P. Hence, the two-stage problem formulation in MPD can be defined as follows:

$$I_o = (1-M) \odot F(I_h, B) + M \odot G(F(I_h), S, P) \tag{9}$$

where F corresponds to the brightness correction that performs pixel-wise dynamic convolution guided by brightness map B. G represents the brightness-missing region recovery process guided by structural semantics S and exposure prompts P.

First, MPD adjusts the brightness of the input image I_h using a Brightness Correction Net η_θ. This network utilizes dynamic convolutions guided by the brightness map B to adaptively adjust the brightness of the input image I_h on a per-pixel basis. This adjustment helps to correct the brightness of the image and also helps to accurately identify brightness-missing regions M. Subsequently, it identifies M in the image by applying a brightness threshold H. We employ a semantic segmentation model to create a semantic map S to guide MPD in recovering brightness-missing regions to ensure that the semantically restored image remains consistent with input images. Throughout the process, MPD adjusts the brightness levels of the generated image using exposure prompts P to ensure that the recovered brightness-missing regions are harmonious with the brightness-corrected image. By integrating these steps into a coherent workflow, our approach effectively tackles the challenges of image restoration under extreme exposure conditions.

C. Multiple Low- and High-level Priors

To restore brightness-missing regions, we utilize the low-level brightness priors, structural semantics priors, and exposure semantics priors to assist in the restoration.

1) Low-level Brightness Priors: The gray-scale value of each pixel in the input color image I_h is obtained to create the brightness map B. On the one hand, the brightness map B is fed into the Brightness Correction Net η_θ to guide the adaptive distribution of convolution kernels in Brightness-aware Dynamic convolution. On the other hand, the brightness map B is used to distinguish brightness-missing regions and obtain the brightness-missing region mask M. We define a brightness threshold H to identify brightness-missing regions in images. For an 8-bit color depth RGB image, the range of its brightness values spans from 0 to 255. To differentiate brightness-missing regions, we set the value of H to 1 for underexposure conditions and 254 for overexposed conditions. Specifically, we consider a region to have brightness missing when the brightness value is below 1 or above 254.

2) High-level Structural Semantics Priors: Different from ControlNet, which only inputs semantic segmentation as a static condition, our method dynamically fuses the brightness-missing region mask M and the semantic map S, introducing a cross-scale semantic attention mechanism during the noise prediction phase of the diffusion model (as shown in the interaction between the semantic encoder τ_θ and the noise prediction network in Fig. 2). This forces the generated content to strictly align with the semantic context of the non-missing regions. Although details may be lost in extreme overexposure or underexposure cases, these regions typically do not cover the entire image, and their surroundings are easily extractable to obtain valuable semantic features. Leveraging this, we employ the semantic map S, generated by the Oneformer [38], a universal image segmentation network, to guide the recovery of content in these regions. Structural Semantics Priors maintains the semantic structure consistency between the recovered and unrecovered images when feeding the brightness correct image I_c to the diffusion model.

3) High-level Exposure Semantics Priors: Exposure semantics prompts P generate dynamic weights through the CLIP text encoder, which are aligned with image features in a cross-modal manner, rather than the fixed edge or segmentation conditions in ControlNet. This design can adaptively adjust the brightness distribution of the generated regions, solving the balance issue of local overexposure/underexposure under extreme exposure. The Structural Semantics Priors provide structural guidance for the image restoration process, yet there remains a possibility that the brightness of the restored image may be abnormal. To address this, we utilize CLIP's [39] text encoder to extract brightness-related textual features and align them with the image features during restoration. This alignment allows us to adjust the brightness of the generated images, thereby preventing any abnormal brightness in the restored images. Specifically, we select exposure prompts P with a positive prompt {Avoid blown-out highlights} and a

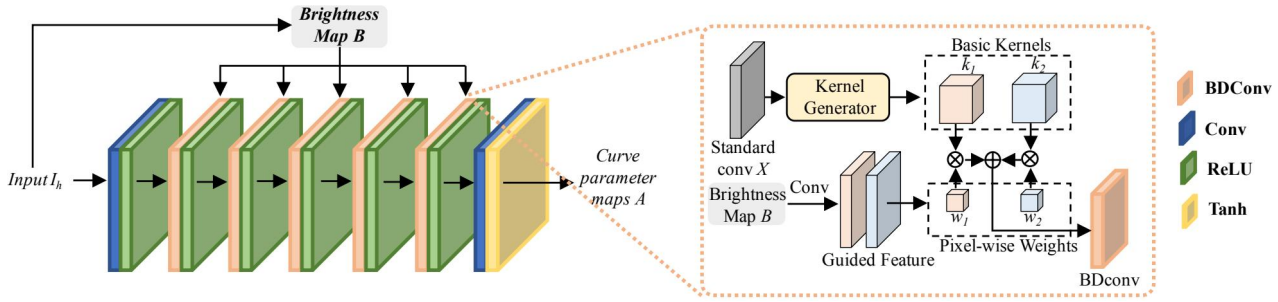


Fig. 3. Illustration of Brightness Correction Net η_θ and Brightness-aware Dynamic convolution. BCN uses seven CNN layers, with ReLU activation in the first six and Tanh in the last. It includes a brightness-aware dynamic convolution in layers two to six to adapt feature extraction to brightness levels.

Negative prompt {Overexposed areas} as exposure semantics priors and input them into Textual Encoder γ_θ to extract brightness-related textual features, which guide the brightness adjustment during the image generation process.

D. Brightness Correction

MPD corrects the brightness of the input images I_h under the guidance of low-level brightness priors. Specifically, we design the Brightness Correction Net η_θ (BCN) to correct the brightness of images with abnormal exposure by predicting per-pixel brightness adjustment parameters A. BCN uses seven CNN layers, with ReLU activation in the first six and Tanh in the last. It includes a brightness-aware dynamic convolution (BDconv) in layers two to six to adapt feature extraction to brightness levels. To minimize information loss, BCN avoids upsampling or downsampling. The structure of BCN is shown in Fig. 3, upon obtaining the pixel-level curve parameters A, inspired by Zero-DCE [1] and [40], the image adjustment parameters follow the iterative adjustment formula:

$$I_c(n) = I_c(n-1) - A * (I_c^2(n-1) - I_c(n-1)) \quad (10)$$

where n is the number of iterations for adjusting the brightness, which is set to 8 in our approach.

Inspired by [41], the Brightness-aware Dynamic convolution (BDconv) is designed to adaptively adjust the correction strength according to the varying exposure conditions. Using the brightness map B extracted from the input image I_h and a standard convolution X, the BDconv kernels are generated by predicting a set of basic kernels and guided features. The structure of BDconv is illustrated in Fig. 3. For the standard convolution X, a kernel generator produces two basic kernels, k_1 and k_2. The brightness map B is input into a convolution layer to predict guided features, which determine pixel-wise weights w_1 and w_2. Under the guidance of these weights, the basic kernels k_1 and k_2 are allocated to regions with different brightness levels, enabling Brightness-aware Dynamic convolution to perform targeted feature extraction. This adaptive approach allows BCN to learn different mapping parameters for different brightness regions in the input image, thereby adaptively optimizing the correction strength.

E. Missing Region Recovery

The Brightness Correction Net η_θ has adjusted the brightness of the input image I_h on a pixel-by-pixel basis to return them to normal light levels and obtain the corrected image I_c:

$$I_c = \eta_\theta(I_h, B) \quad (11)$$

However, for the brightness-missing regions, the BCN η_θ has little effect due to the loss of brightness information, so in this section, we will solve this problem.

After obtaining I_c, the Image Encoder E processes the image I_c to produce a latent feature z_{input} t through:

$$z_{input} = E(I_c) \quad (12)$$

which is then subjected to random noise to create a noisy latent feature z_T in the diffusion model (where T represents the total time step). The noisy latent feature is generated by:

$$z_T = \alpha_T z_{input} + \sigma_T \epsilon \quad (13)$$

where α_T and σ_T are predefined hyperparameters controlling the noise schedule. To guarantee the restored brightness-missing regions in the image align semantically with the original features, we feed the semantic map S into the Semantic Encoder τ_θ to derive the semantic condition c. This condition then directs the Noise Prediction Network ϵ_θ to predict the noise added to the latent feature z_T. Furthermore, the CLIP text-encoder γ_θ extracts a text-guided condition y, linking the exposure prompts P with the generated image's brightness. This dual guidance ensures semantic accuracy and visual quality in the restored content.

The structural semantics condition c and exposure semantics condition y implement synergistic control over the diffusion process through distinct mechanisms. Specifically: The semantic map S is encoded by ControlNet [32] to generate spatially aligned feature maps. These features are injected into the UNet's intermediate layers via zero-initialized convolution layers, establishing pixel-wise correspondence between the semantic structure and the generated content. This process can be formulated as:

$$h^{(l)}(z_t) = h^{(l)}_{ori}(z_t) + \tau^{(l)}_\theta(S) \quad (14)$$

where $h^{(l)}_{ori}$ denotes the original feature at layer l, $\tau^{(l)}_\theta$ represents the ControlNet branch with zero-initialized convolutions that inject semantic map S, enforcing structural consistency

through residual feature modulation. Structural semantics condition $c = \tau_\theta(S)$. The CLIP text encoder γ_θ projects exposure prompts P into a joint embedding space, establishing cross-modal alignment between textual semantics and visual features. This alignment is achieved through cross-attention layers in the diffusion model:

$$\text{Attention}(Q^{(l)}, K_y^{(l)}, V_y^{(l)}) = \text{softmax}\left(\frac{Q^{(l)}(K_y^{(l)})^T}{\sqrt{d}}\right) V_y^{(l)} \tag{15}$$

where query $Q^{(l)}$ comes from visual features, key $K_y^{(l)}$ and value $V_y^{(l)}$ are derived from text-guided condition $y = \gamma_\theta(P)$. Here d denotes the feature dimension.

The joint guidance (c, y) creates a dual-control system: structural semantics maintain geometric and contextual coherence through spatial feature constraints, while exposure semantics regulate global illumination distribution via cross-modal semantic alignment. The guided noise prediction is represented as:

$$\hat{\epsilon}_\theta\left(z_t, t, (c, y)\right) = \epsilon_\theta\left(z_t, t, \emptyset\right) + k\left(\epsilon_\theta\left(z_t, t, (c, y)\right) - \epsilon_\theta\left(z_t, t, \emptyset\right)\right) \tag{16}$$

where k is a factor adjusting the strength of the guidance.

Following noise prediction, the Denoising Diffusion Implicit Models (DDIM) [37] sampling scheduler, as shown in Eq. 8, is used to accelerate the sampling process and gradually denoise to obtain z_0:

$$z_{t-1} = \alpha_{t-1} z_0 + \sigma_{t-1}\left(\zeta\epsilon + \sqrt{1-\eta^2}\hat{\epsilon}_\theta(z_t, t, (c, y))\right), \quad \epsilon \sim \mathcal{N}(0, I) \tag{17}$$

where ζ is typically set to 0, α and σ are predefined hyperparameters. To finalize the restoration, we pass the predicted z_0 through Image Decoder D to generate the image $\hat{I}_o = D(z_0)$. The final image is then composed using the following formula:

$$I_o = (1 - M) \odot I_c + M \odot \hat{I}_o \tag{18}$$

The whole algorithm is shown in Algorithm 1.

Algorithm 1 MPD.

Input: Input image, I_h; Brightness Correction Net, η_θ; Semantic Segmentation Net, δ_θ; Semantic ControlNet, τ_θ; Exposure Prompt, P; Textual Encoder, γ_θ; Noise Prediction Network ϵ_θ, Image Encoder, E; Image Decoder, D;
Output: Image I_o after MPD;

Extract B based on I_h and obtain Brightness-missing Region M;
$I_c = \eta_\theta(I_h, B)$;
$c = \tau_\theta(\delta_\theta(I_c))$, $y = \gamma_\theta(P)$;
$z_{input} = E(I_c)$;
$\epsilon \sim \mathcal{N}(0, \mathbf{I})$;
$z_T = \alpha_T z_{input} + \sigma_T \epsilon$;
for all t from T to 0 do:
  $\hat{\epsilon}_\theta\left(z_t, t, (c, y)\right) = \epsilon_\theta\left(z_t, t, \emptyset\right) + k\left(\epsilon_\theta\left(z_t, t, (c, y)\right) - \epsilon_\theta\left(z_t, t, \emptyset\right)\right)$;
  $z_{t-1} = \alpha_{t-1} z_0 + \sigma_{t-1}(\zeta\epsilon + \sqrt{1-\eta^2}\hat{\epsilon}_\theta\left(z_t, t, (c, y)\right))$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$;
end for;
$\hat{I}_o = D(z_0)$;
$I_o = \hat{I}_o \odot M + I_c \odot (1 - M)$;
return I_o;

F. Loss Functions

In our approach, the subnetworks, including Semantic Segmentation, Semantic Encoder, Textual Encoder, Image Encoder, and Image Decoder, directly leverage the trained models, and their weights are frozen. We train the Brightness Correction Net η_θ and the Noise Prediction Network ϵ_θ. For the BCN, we utilize an exposure control loss [1], denoted as L_{exp}, which measures the deviation between the average brightness of a local region and the desired brightness for proper exposure. Additionally, to assist the network in accurately estimating the illumination and parameter maps, we apply a local smoothness term [42], denoted as L_{tv}, during the brightness correction process. To minimize color cast and ensure that the RGB channels' distribution remains close, we employ a color constancy loss [43], denoted as L_{col}. Furthermore, we use L_{spa} [1] to measure changes in pixel value differences between adjacent regions post-correction, aiming to preserve spatial consistency. Therefore, the brightness correction loss L_{bc} is formulated as:

$$\mathcal{L}_{\text{bc}} = \lambda_1 \mathcal{L}_{\text{exp}} + \lambda_2 \mathcal{L}_{\text{tv}} + \lambda_3 \mathcal{L}_{\text{col}} + \lambda_4 \mathcal{L}_{\text{spa}} \tag{19}$$

where λ_1, λ_2, λ_3 and λ_4 are the balancing hyperparameters.

To minimize the occurrence of brightness anomalies in the images generated by the Stable Diffusion model, we fine-tune the Noise Prediction Network ϵ_θ using the following mean square loss:

$$\mathcal{L}_{br} = \mathbb{E}_{z, \epsilon\sim\mathcal{N}(0,\mathbf{I}), t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t\right)\|_2^2\right] \tag{20}$$

where ϵ_θ is trained to predict the noise ϵ contained in the input z_t at any time t.

G. Training

1) Training Set Preparation: The training pipeline employs a two-stage data strategy to accommodate the training of the brightness correction net and the fine-tuning of the noise prediction network. For the initial training of BCN, we need a large number of images to learn how to map underexposed/overexposed images to normal exposure. Since our brightness correction net's training process is unsupervised and data-insensitive, we choose images from the MS-COCO dataset [44] and apply random brightness adjustments to simulate underexposed/overexposed conditions. This selection is motivated by the dataset's extensive coverage of standard illumination conditions and strict exclusion of images overlapping with our test benchmarks (SICE and Adobe FiveK). To simulate extreme exposure variations (±3 EV) within sRGB constraints, we utilize the Adobe Camera Raw SDK in Photoshop for controlled brightness manipulation.

TABLE I
PSNR ↑, SSIM ↑, UNIQUE (UN.) ↑, NIQE ↓, LPIPS ↓, CLIP-IQA ↑, U.S. ↑ SCORES ON SICE AND E-FIVEK DATASETS UNDER EXTREME UNDEREXPOSURE. THE BOLDFACE AND UNDERLINED ENTRIES DENOTE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

Methods	SICE							E-FiveK						
	PSNR	SSIM	UNIQUE	NIQE	LPIPS	CLIP-IQA	U.S.	PSNR	SSIM	UNIQUE	NIQE	LPIPS	CLIP-IQA	U.S.
Input	7.2900	0.1409	0.1173	5.9352	0.6844	0.4615	2.3612	9.8088	0.2875	-0.0529	6.3456	0.4428	0.3823	2.4869
HDRCNN [22]	6.2869	0.0555	-0.3204	8.6555	0.8367	0.3977	1.3968	7.8016	0.1083	-0.5270	8.8706	0.6772	0.3541	1.6858
Zero-DCE [1]	10.7783	0.4093	0.5527	4.7390	0.4482	0.4376	3.1558	17.2077	0.6195	0.4616	5.4369	0.3043	0.3895	3.2547
CMEC [46]	8.9377	0.2731	0.3635	6.0113	0.6252	0.3860	2.1547	18.9343	0.6256	0.2344	6.2352	0.3303	0.3609	1.9314
MSEC [6]	10.3130	0.3494	0.2760	5.1044	0.5995	0.2814	2.0385	18.5191	0.6256	0.0191	6.0567	0.3926	0.2966	2.8562
FECNet [4]	11.3242	0.4232	0.1129	3.9022	0.4811	0.3147	3.2569	19.8765	0.6781	0.0744	5.0928	0.3081	0.3161	3.1569
LCDPNet [5]	11.5386	0.4589	0.5157	4.5025	0.4279	0.3861	2.5541	17.4115	0.6195	0.3735	5.8892	0.3090	0.3819	2.6314
LANet [47]	11.2123	0.5551	0.4453	4.5454	0.4509	0.3876	3.2147	**20.3520**	**0.7639**	0.5417	5.4087	0.3006	0.3957	3.3423
Retinexformer [15]	8.5441	0.2390	0.4773	6.0245	0.6051	0.4489	2.2559	16.6373	0.5576	0.4952	5.8452	0.3223	0.4082	2.6847
RetinexMamba [16]	11.2837	0.4510	0.5873	3.5671	0.4722	0.3758	2.6320	15.9237	0.7064	0.5401	5.1505	0.3104	0.3567	3.0441
CoTF [17]	12.7222	0.5544	0.3558	4.1986	**0.4121**	0.3351	2.8132	17.7618	0.7367	0.4097	5.4992	0.2978	0.3529	2.9644
CSEC [18]	10.7060	0.4216	0.2556	4.3583	0.4498	0.3588	2.9655	14.7625	0.6113	0.3225	5.6601	0.3258	0.3711	3.1124
UNICE [48]	**13.4730**	**0.5866**	0.4883	4.4950	0.4393	0.4025	3.2216	17.6665	0.7066	0.6566	5.2393	0.3018	0.4758	3.3418
LLDiffusion [49]	12.6869	0.5196	0.5600	3.8407	0.4260	0.4963	3.0012	17.6297	0.6981	0.5399	5.4233	**0.2837**	**0.5078**	3.2064
Ours	12.6486	0.5669	**0.8215**	**3.4397**	0.4192	**0.5251**	**3.4672**	17.6859	0.6792	**0.8279**	**4.3402**	0.2955	0.4861	**3.5149**

The subsequent fine-tuning stage on the SICE dataset addresses two critical considerations. First, SICE includes images with varying exposures from multiple real-world scenarios and their corresponding high-quality ground truth, enabling supervised calibration of the diffusion model's noise prediction parameters. Second, the inherent data efficiency of Low-Rank Adaptation (LoRA) [45] allows effective model specialization using only 20 carefully selected samples, circumventing the need for large-scale annotated data. LoRA freezes the pre-trained model weights and injects trainable low-rank matrices into specific layers, enabling adaptation with minimal computational overhead and reduced risk of overfitting. This approach is particularly suitable for scenarios with scarce training data, as it constrains the fine-tuning process to a low-dimensional subspace of the original parameter space.

2) Training Details: For training BCN, we employ an end-to-end approach, resizing images to 512 × 512. All experiments are conducted on a single NVIDIA 3090 GPU.Network biases are initialized to a constant value, and we optimize using the Adam optimizer [50] with a fixed learning rate of 1×10^{-4}, a batch size of 4, training for 200 epochs. The loss function weights are set as follows: $\lambda_1 = 10, \lambda_2 = 200, \lambda_3 = 5$, and $\lambda_4 = 1$. We use 20 images of normal exposure from the SICE dataset to apply a low-rank approximation scheme [45] to fine-tune the Noise Prediction Network ϵ_θ to minimize the objective described in Eq. 20. These images are selected to cover diverse indoor and outdoor scenes under normal illumination, and they have no overlap with the training or test sets used in this work. During fine-tuning, we inject LoRA matrices into the cross-attention layers of the pre-trained Stable Diffusion v1.5 [14]. We use the Adam optimizer [50] with a learning rate of 1e-5 and a batch size of 1. The small batch size helps reduce gradient noise, while the low learning rate ensures stable convergence. In addition, we utilize the pre-trained Oneformer [38] on the ADE20K [51] dataset for generating semantic segmentation maps, incorporate the pre-trained Semantic Encoder [32] for semantic feature guidance, and employ the pre-trained Stable Diffusion v1.5 [14]. For diffusion inference, we set the CFG scale to 7, the denoising strength to 0.76, and fix the random seed for reproducibility. The sampling strategy of DDIM is adopted, with the number of sampling steps set to 20.

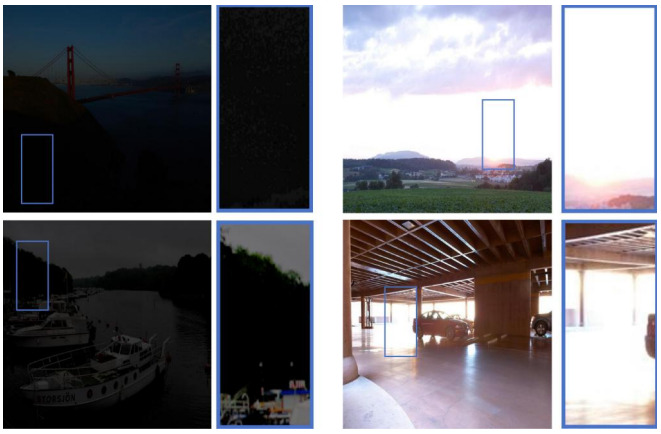


Fig. 4. The inputs from SICE dataset and E-FiveK dataset. The two images are extremely underexposed and overexposed images respectively, both of which involve brightness-missing regions.

IV. EXPERIMENTAL RESULTS

A. Experiment Setting

Regarding the testing sets, a few images in existing datasets include extreme exposure conditions. The images in our testing sets are sourced from the SICE dataset [52] and the Adobe FiveK dataset [53]. The SICE dataset comprises multiple

TABLE II
PSNR ↑, SSIM ↑, UNIQUE (UN.) ↑, NIQE ↓, LPIPS ↓, CLIP-IQA ↑, U.S. ↑ SCORES ON SICE AND E-FIVEK DATASETS UNDER EXTREME **OVER**EXPOSURE. THE BOLDFACE AND UNDERLINED ENTRIES DENOTE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

Methods	SICE							E-FiveK						
	PSNR	SSIM	UNIQUE	NIQE	LPIPS	CLIP-IQA	U.S.	PSNR	SSIM	UNIQUE	NIQE	LPIPS	CLIP-IQA	U.S.
Input	11.5943	0.7227	1.1556	3.5979	0.2594	0.4488	1.9631	11.2011	0.6959	0.6525	5.3044	0.2646	0.4007	1.2356
HDRCNN [22]	11.5221	0.5612	0.5687	3.6938	0.3293	0.4520	1.9985	13.4323	0.7105	0.0625	4.9030	0.2887	0.3468	1.1356
Zero-DCE [1]	9.3312	0.6390	0.7898	3.7496	0.3235	0.4510	2.6532	7.6781	0.6072	0.4865	5.2602	0.3618	0.3578	2.4651
CMEC [46]	15.4544	0.7151	**1.3753**	3.3123	0.2335	0.4863	2.1321	17.1493	0.8291	0.7265	4.6492	0.2511	0.4177	2.3654
MSEC [6]	15.3911	0.6821	1.3260	3.5237	0.2438	0.4105	2.3651	16.7227	0.7546	0.6715	4.8313	0.2811	0.4003	2.3261
FECNet [4]	15.2507	0.5878	0.7736	3.2606	0.3077	0.4315	3.1256	18.5349	0.7209	0.0431	4.7465	0.2974	0.3343	3.2659
LCDPNet [5]	14.3432	0.6974	0.9202	3.8446	0.2639	0.4885	2.8641	13.4747	0.6774	0.3853	4.6837	0.2924	0.3393	2.9532
LANet [47]	15.0932	0.6876	0.8305	4.8012	0.2916	0.4846	3.3621	15.0492	0.7123	0.5212	4.9213	0.2620	0.4272	3.2156
Retinexformer [15]	15.1936	0.6983	1.0868	3.3392	0.2581	0.4313	2.7347	14.3945	0.7583	**0.9157**	4.5434	0.2505	0.4109	2.5524
RetinexMamba [16]	15.8601	0.7104	1.1242	3.2772	0.2501	0.4556	2.7681	**18.9666**	**0.8413**	0.5406	4.5668	**0.2169**	0.3857	3.0149
CoTF [17]	16.7123	0.7201	0.8056	3.4881	0.2558	0.3778	2.8961	14.8395	0.7258	0.4862	4.7283	0.2900	0.3412	2.8774
CSEC [18]	15.1202	0.7211	0.9755	3.3380	0.2292	0.4839	3.1653	14.1966	0.6953	0.2815	4.6714	0.2980	0.3434	3.2651
UNICE [48]	16.0981	0.6873	1.0394	3.3894	0.2367	0.4804	3.4214	16.6077	0.7781	0.8749	4.7619	0.2466	0.4753	3.2167
LLDiffusion [49]	11.4190	0.6939	1.0351	3.5054	0.2761	0.4568	2.7457	9.1447	0.6263	0.5555	4.6685	0.3106	0.4511	2.6684
Ours	**17.3639**	**0.7432**	1.2907	**3.2183**	**0.2168**	**0.5117**	**3.6327**	16.8553	0.7792	0.8905	**4.4090**	0.2418	**0.4909**	**3.5426**

groups of images with varying exposure conditions. We select 120 groups and choose the images with the highest and lowest brightness values from each group to represent overexposure/underexposure conditions. The Adobe FiveK dataset contains raw-format images along with corresponding expert-annotated ground truth. We randomly select 140 images from this dataset and perform brightness adjustments using the same method as in the training set, specifically leveraging the Adobe Camera Raw SDK in Photoshop. The adjustment parameters are set to a relative EV of ±3. We designate these images from the Adobe FiveK dataset captured under extreme exposure conditions as the E-FiveK database. We use the SICE and the E-FiveK datasets as our testing sets. It is important to note that our training set and testing set do not intersect.

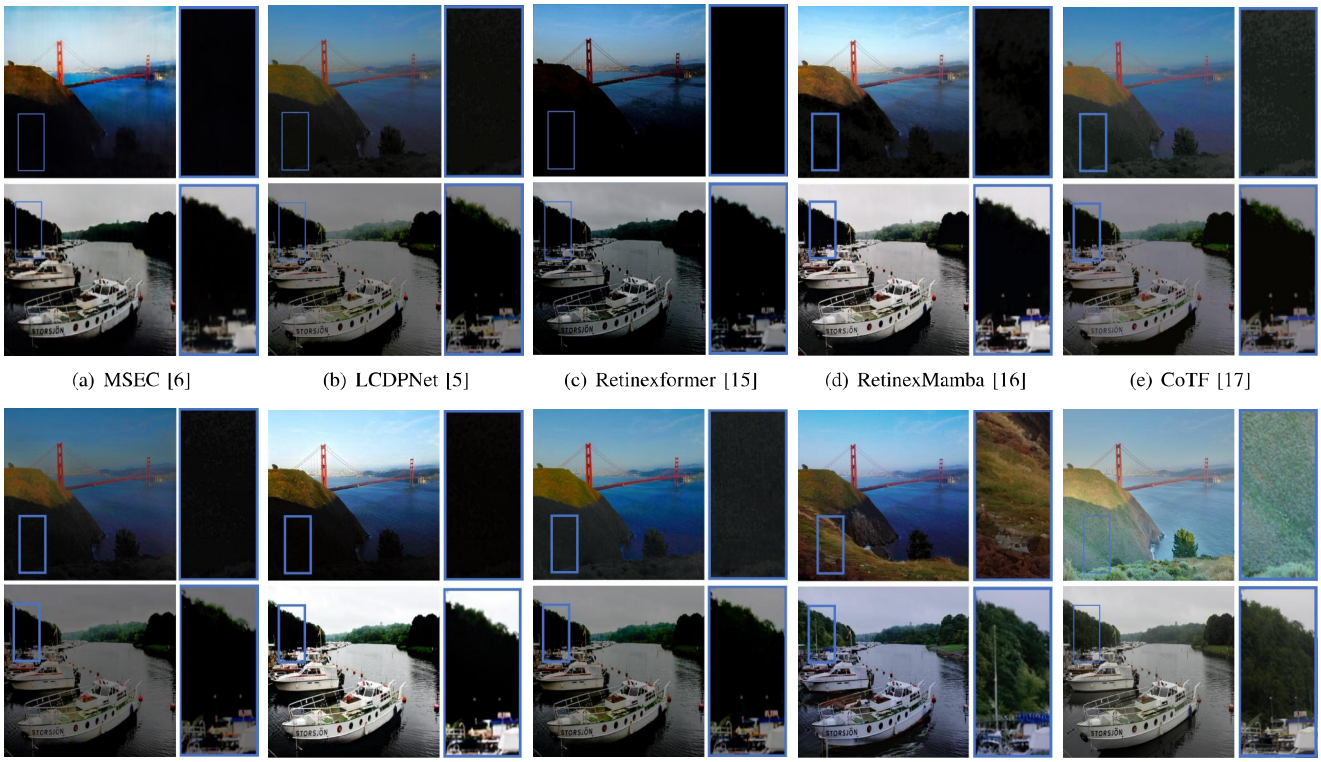
B. Comparisons with State-of-the-arts

1) Quantitative Comparison: To demonstrate the effectiveness of our method, we compare it with representative exposure correction approaches, including MSEC [6], LCDP-Net [5], Retinexformer [15], RetinexMamba [16], CoTF [17], CSEC [18], UNICE [48], and LLDiffusion [49]. Notably, UNICE and LLDiffusion are diffusion-based generative methods that restore images via iterative sampling with strong priors, providing a particularly relevant and challenging comparison in terms of perceptual realism and detail recovery. We evaluate the restored results on the SICE dataset and E-FiveK dataset using both no-reference and full-reference image quality assessments. For no-reference evaluation, we adopt the Natural Image Quality Evaluator (NIQE) [54] and UNIQUE [55] metrics. For full-reference evaluation, we use PSNR (dB) and SSIM as distortion-oriented metrics, and further incorporate LPIPS [56] and CLIP-IQA [57] as perceptual measures: LPIPS quantifies perceptual similarity between the restored image and the reference in a deep feature space, while CLIP-IQA leverages CLIP-based semantic representations to assess overall visual quality, aligning better with human subjective preference. We conduct experiments under two extreme conditions separately: extreme underexposure and extreme overexposure.

As shown in Table I and Table II, our method achieves competitive overall performance on both SICE and E-FiveK, and shows clear advantages on perceptual-related criteria, indicating improved naturalness, visual consistency, and subjective quality. We note that PSNR and SSIM are distortion-based metrics that penalize any pixel-wise deviation from the ground truth, even when such deviations are perceptually plausible or visually preferable. For example, in brightness-missing regions, the input image may have completely lost surface texture information; consequently, without reliable texture cues as a reference, the restored image can present differences in texture and structure compared to the ground truth.

2) Visual Quality Comparisons: In this section, we conduct a detailed examination of MPD in terms of brightness, color, contrast, and naturalness under extreme exposure conditions. As shown in Fig. 5 and Fig. 6, in both underexposure and overexposure scenarios, our method makes global adjustments to the image's brightness and delicately restores the color and contrast of the brightness-missing regions. Our method also demonstrates a natural and harmonious restoration of brightness-missing regions. Overall, our method achieves satisfactory results in terms of brightness, color, and naturalness.

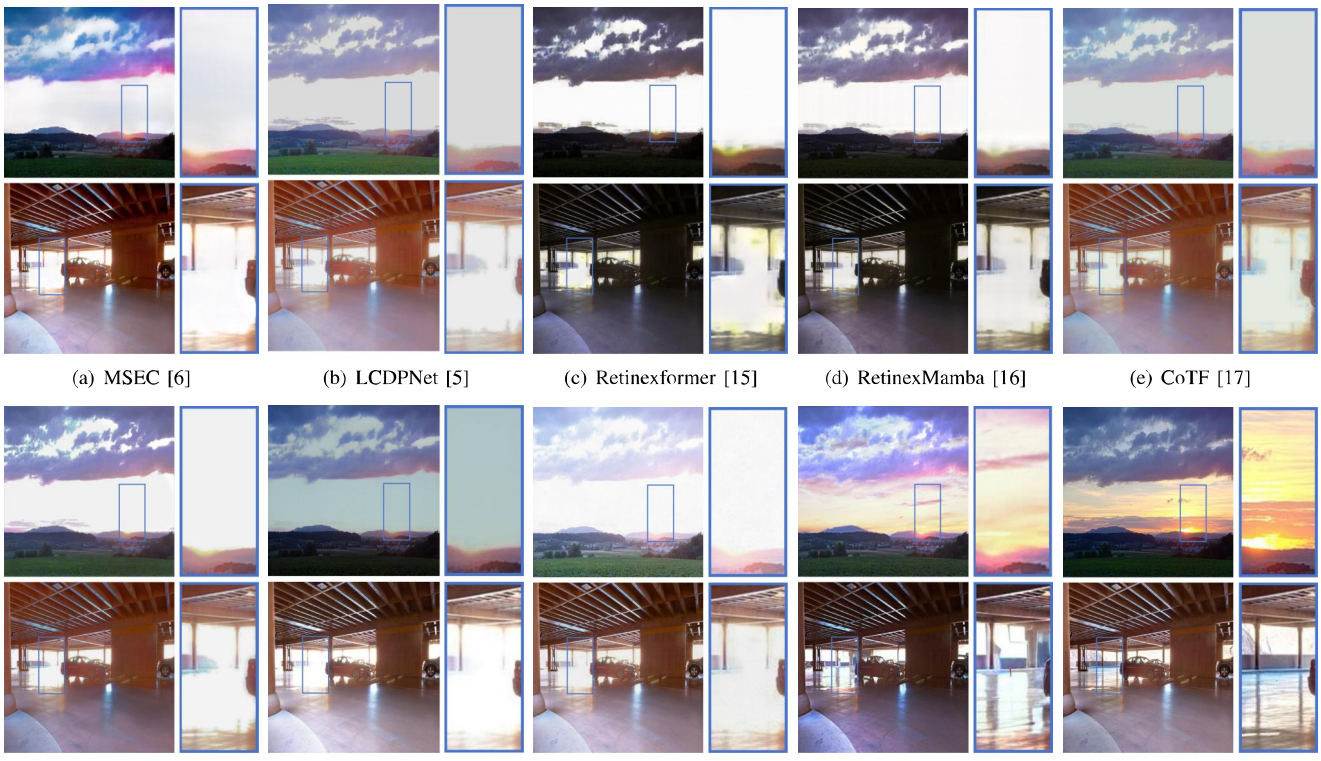
3) Human subjective survey: We conduct a human subjective survey for comparison. For the restored images in the testing set produced by nine different methods, we invited 15 human subjects, who were asked to rank the enhanced images. These subjects are instructed to consider the following criteria: i) whether the image brightness, color, and contrast are normal; ii) whether there are noticeable artifacts of overexposure or



(a) MSEC [6] (b) LCDPNet [5] (c) Retinexformer [15] (d) RetinexMamba [16] (e) CoTF [17]

(f) CSEC [18] (g) UNICE [48] (h) LLDiffusion [49] (i) Ours (j) GT

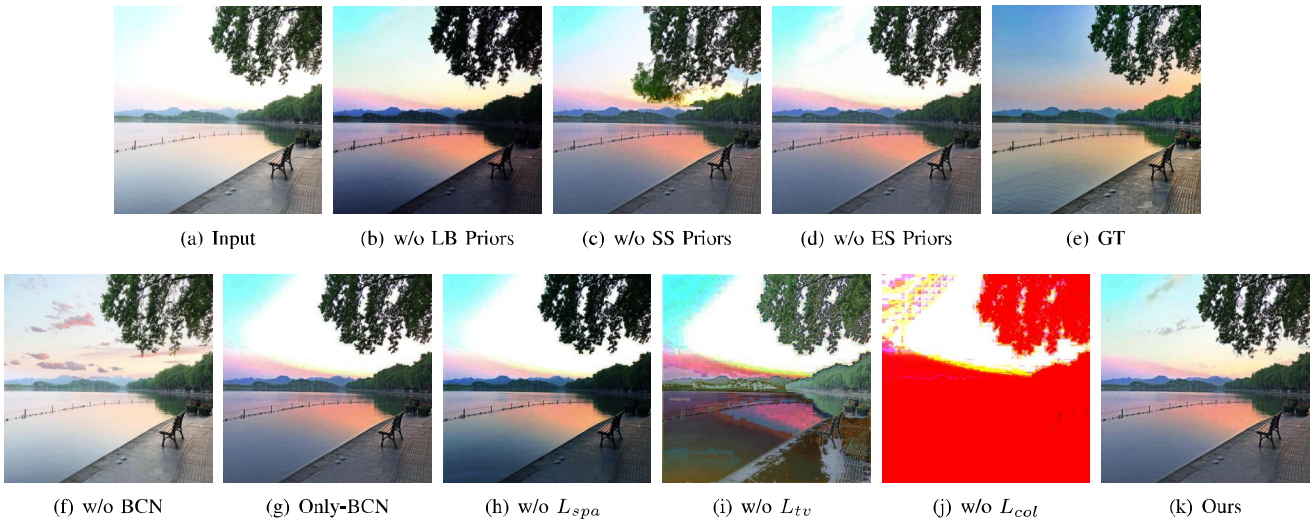
Fig. 5. Visual comparison of **extremely underexposed** images from the SICE dataset and E-FiveK dataset. Compared with peer methods, our restoration results achieve more satisfactory outcomes in terms of brightness, color, and naturalness.



(a) MSEC [6] (b) LCDPNet [5] (c) Retinexformer [15] (d) RetinexMamba [16] (e) CoTF [17]

(f) CSEC [18] (g) UNICE [48] (h) LLDiffusion [49] (i) Ours (j) GT

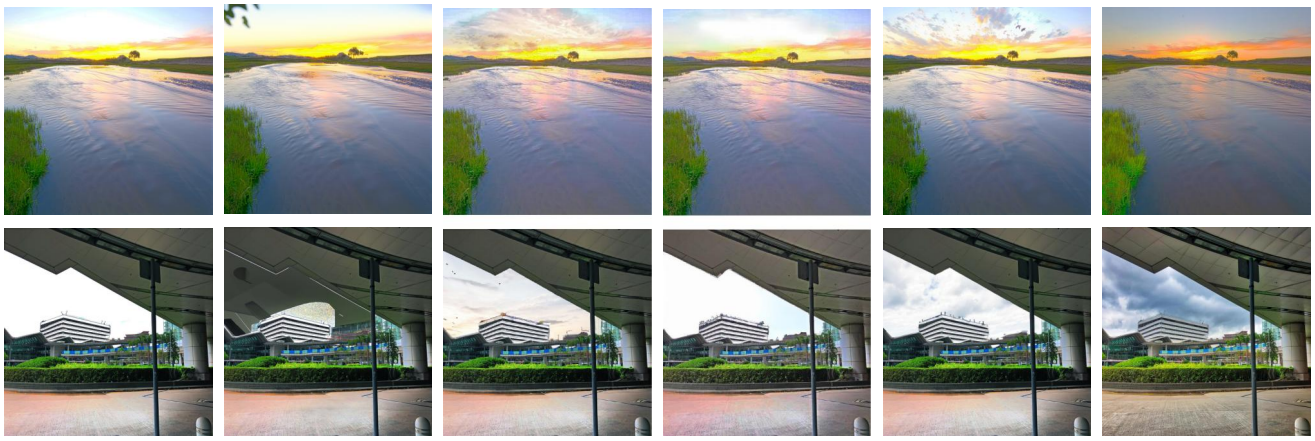
Fig. 6. Visual comparison of **extremely overexposed** images from the SICE dataset and E-FiveK dataset. Compared with peer methods, our restoration results deliver more pleasing performance in terms of brightness, color fidelity, and naturalness.



(a) Input (b) w/o LB Priors (c) w/o SS Priors (d) w/o ES Priors (e) GT

(f) w/o BCN (g) Only-BCN (h) w/o L_{spa} (i) w/o L_{tv} (j) w/o L_{col} (k) Ours

Fig. 7. Ablation study of our framework and loss functions. (b) - (f) demonstrating the effectiveness of the prior knowledge and network structure we introduced, and (g) - (i) reflecting the roles of various loss functions.



(a) Input (I_c) (b) Repaint [26] (c) Uni-paint [27] (d) DiffBIR [24] (e) Ours (f) GT

Fig. 8. Display of brightness-missing regions recovery without semantic conditions. Our method produces accurate content and maintains appropriate brightness, contrast, and naturalness, resulting in satisfactory outcomes.

TABLE III
ABLATION STUDY OF OUR FRAMEWORK AND LOSS FUNCTIONS.

Methods	NIQE ↓	UNIQUE↑	PSNR ↑	SSIM ↑
Input	3.7209	1.1253	11.5740	0.7077
w/o LB Priors	3.7375	0.9738	13.8971	0.6446
w/o SS Priors	3.3656	1.3652	15.2321	0.5611
w/o ES Priors	3.2652	1.3236	14.4522	0.6235
w/o BCN	3.1659	1.3876	12.2485	0.6257
Only-BCN	3.5623	1.1235	14.2362	**0.7166**
w/o L_{spa}	3.5656	0.9889	14.3015	0.6523
w/o L_{tv}	4.9376	-0.2685	11.6619	0.4364
w/o L_{col}	13.9696	-1.9583	6.8382	0.2092
Ours	**3.0065**	**1.3975**	**15.4842**	0.6776

underexposure; iii) whether the image exhibits overall or partial visual inconsistencies. We specify a scoring range from 1 to 5 for each image, with a higher value indicating better image quality. As shown in Table I (U.S.), our approach achieves the best results.

C. Ablation Study

1) Investigation of the framework: To demonstrate the effectiveness of the general configuration of our method, we perform several ablation settings and present the results in Table III. Specifically:

i) w/o Low-level Brightness Priors (w/o LB Priors): Do not use low-level brightness priors and replace brightness-aware dynamic convolution with regular convolution in the Brightness Correction Net η_θ;

ii) w/o Structural Semantics Priors (w/o SS Priors): Do not use structural semantics priors and remove Semantic Encoder τ_θ from MPD;

iii) w/o Exposure Semantics Priors (w/o ES Priors): Do not use exposure semantics priors and remove Textual Encoder γ_θ from MPD;

iv) w/o BCN: Remove the Brightness Correction Net η_θ from the MPD, and the image will not go through the BCN;

v) Only BCN: Images are processed only through the Brightness Correction Net η_θ.

As shown in Table III, the results from the second to fourth lines indicate that the three types of prior knowledge we introduced are essential and play an important guiding role in image restoration. The results of w/o BCN demonstrate the essential role of brightness correction in BCN. The comparison between W/o LB Priors and Only BCN results also confirms the effectiveness of BDconv.

The visual results of the ablation study are shown in Fig. 7. The results w/o LB Priors exhibit a darker foreground, indicating that this portion of brightness has been overcorrected, highlighting the effectiveness of low-level brightness priors and dynamic convolutions. In the results w/o SS Priors, some semantic inconsistency may occur due to the lack of structural-semantic information constraints, such as in the tree part of the image. Results w/o ES Priors show that, due to the lack of exposure semantic information, although the sky portion has recovered texture information, the brightness remains too bright. The image without BCN is generally brighter, lacking global brightness correction. While Only BCN performs brightness correction, the texture in brightness-missing regions still fails to recover.

2) Ablation study of loss functions: We also conduct an ablation study on the loss functions in Table III. By comparing the different loss functions listed in the table, we find that all of them improve the overall performance of our method. L_{spa} term preserves the spatial consistency of adjacent regions between the input image and the corrected image, while L_{tv} promotes image smoothness and prevents excessive artifacts and false contours at the edges. L_{col} maintains color channel balance and avoids color deviations.

3) Ablation study of CLIP Prompt: To evaluate the model's sensitivity to prompt semantics, we design four prompt sets for both extreme low-light and extreme high-light scenarios, including default, shifted, conflicting, and weakly related prompts, and conduct controlled comparisons with all other settings fixed. The results show that the default prompts consistently achieve the best performance, while the other three settings lead to varying degrees of degradation, with the conflicting prompts causing the most severe drop. This indicates that the model is clearly sensitive to the semantic direction of the prompts, and prompts aligned with the restoration objective are the most effective for recovery.

D. Comparative Study with Diffusion-based Methods

To further demonstrate the effectiveness of our method, we send the corrected image I_c obtained through Brightness Correction Net to other diffusion-based restoration methods [24], [26], [27] for recovering. As shown in Fig. 8, the semantics of generated content may exhibit strong randomness due to the lack of semantic features for content control. This random content is interfered with by the texture of the edges of surrounding objects and may be distorted as a result. In contrast, our method produces accurate content and maintains appropriate brightness, contrast, and naturalness, resulting in satisfactory outcomes. We compare our results with these methods on both no-referenced and full-referenced quality assessments. As shown in Table V, our method achieves the best results on all four metrics.

E. Semantic Coherency

We first visualize the brightness map B guiding BDconv to examine the rationality of the guiding conditions. As shown in Fig. 10, our method effectively grades the luminance to apply different dynamic convolution kernels. To further demonstrate the validity of semantic features for guiding content recovery in brightness-missing regions, we show I_c and Ground-truth semantic segmentation in scenarios with brightness-missing regions. As shown in Fig. 11, the semantic features identified from I_c align with the ground truth, even the texture of the brightness-missing regions is completely lost.

V. Discussion

A. Computational Complexity and Practical Considerations

While MPD, as a diffusion-based model, inherits the higher computational cost typical of high-performance generative approaches for restoring ill-posed brightness-missing regions, its design incorporates optimizations to manage cost effectively for the challenging task:

1. As detailed in Table V, in the two-stage design, the lightweight Brightness Correction Net (BCN, 0.11M parameters) first performs a single-pass global adjustment. The computationally intensive diffusion process is then applied only to the corrupted regions specified by a binary mask, rather than to the entire image, significantly reducing the diffusion denoising burden.

For fine-tuning, the pre-trained diffusion prior is fine-tuned using Low-Rank Adaptation (LoRA), updating only a minimal subset of parameters.

For inference, on an NVIDIA RTX 3090 GPU, inference for a 512×512 image takes 4.17 seconds, which is favorable compared to the state-of-the-art diffusion-based restoration models (e.g., RePaint) that often require similar or more steps.

2. MPD justified in quality-critical applications, where restoring severely degraded content is paramount. For real-time or mobile constraints, the current model is well-suited for a cloud-based inference service. The lightweight BCN can run on-device for initial correction, balancing bandwidth and latency. Currently, the deployment of MPD on consumer devices (e.g., smartphones) with a cloud-based service still cannot achieve real-time inference. However, many practical scenarios for extreme exposure restoration do not demand real-time processing (users may wait seconds to salvage a crucial photo, or batches can be processed in the cloud). With anticipated advancements in hardware and ongoing model lightweighting techniques, real-time restoration on mobile devices becomes a feasible future goal.

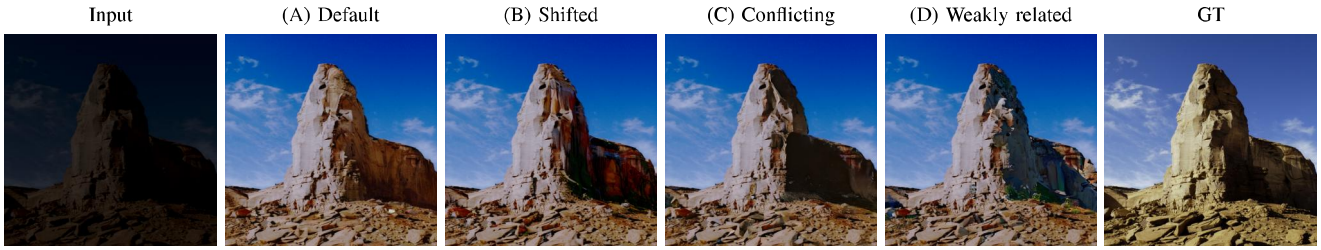
In summary, MPD establishes a quality-efficiency trade-off through deliberate design, offering a practical, high-quality solution for restoring the ill-posed brightness-missing regions.

B. Failure Case

As shown in Fig. 12, when the exterior walls of a building turn white due to overexposure or reflection, they are mistakenly identified for recovering as brightness-missing regions. The current brightness threshold-based detection mechanism may misclassify naturally bright surfaces (e.g., white exterior

TABLE IV
SENSITIVITY ANALYSIS OF PROMPT SETTINGS FOR UNDEREXPOSED AND OVEREXPOSED CASES ON THE SICE AND E-FIVEK DATASETS, WHERE (A)–(D) DENOTE DEFAULT, SHIFTED, CONFLICTING, AND WEAKLY RELATED PROMPTS.

	Prompt Settings	SICE				E-FiveK			
		PSNR↑	SSIM↑	UNIQUE↑	NIQE↓	PSNR↑	SSIM↑	UNIQUE↑	NIQE↓
Underexposed	**(A)** + Recover natural brightness - Underexposed regions	**12.6486**	**0.5669**	**0.8215**	**3.4397**	**17.6859**	**0.6792**	0.8279	**4.3402**
	(B) + Enhance dramatic contrast - Soft dark tones	11.7768	0.4704	0.8115	3.5045	17.4850	0.6641	**0.8453**	4.4207
	(C) + Preserve deep darkness - Bright appearance	11.7752	0.4725	0.8158	3.5146	17.5198	0.6485	0.8231	4.4426
	(D) + Enhance metallic texture - Smooth surfaces	11.6777	0.4655	0.8031	3.5168	17.5043	0.6497	0.8152	4.4491
Overexposed	**(A)** + Recover natural highlights - Overexposed regions	**17.3639**	**0.7432**	**1.2907**	**3.2183**	**16.8553**	**0.7792**	**0.8905**	**4.4090**
	(B) + Enhance glossy reflections - Soft highlight transitions	16.9597	0.7282	1.2781	3.2354	16.0909	0.7702	0.8767	4.5063
	(C) + Preserve intense glare - Natural shading	16.9562	0.7294	1.2657	3.2549	16.1106	0.7717	0.8784	4.4845
	(D) + Enhance sharp edges - Soft boundaries	16.9492	0.7311	1.2490	3.3041	16.0972	0.7735	0.8690	4.4812



(a) Visual comparison under extreme underexposure restoration.

(b) Visual comparison under extreme overexposure restoration.

Fig. 9. Visual comparison of different prompt settings under extreme underexposure and overexposure conditions.

TABLE V
EXPERIMENTAL ON DIFFUSION-BASED RECOVERY METHODS.

Methods	PSNR↑	SSIM↑	UNIQUE↑	NIQE↓	LPIPS↓	CLIP-IQA↑
Input(I_c)	15.7396	0.7166	1.1235	3.5623	0.1996	0.5372
Repaint [26]	15.2124	0.6255	1.2588	3.2416	0.2601	0.5095
Uni-paint [27]	16.5542	0.6917	1.2775	3.2612	0.2387	0.4936
DiffBIR [24]	16.8362	0.7248	1.2791	3.2219	0.2303	0.5012
Ours	**17.3639**	**0.7432**	**1.2907**	**3.2183**	**0.2168**	**0.5117**

TABLE VI
THE PARAMETERS (M) IN EACH SUBNETWORK OF MPD.

Subnetwork	Params. (M)
Brightness Correction Net	0.11
Noise Prediction Network	859.52
Image Encoder	34.16
Image Decoder	123.13
Semantic Segmentation Network	219.0
Semantic ControlNet	429.60

walls or marble facades) as overexposed regions. This occurs because the fixed brightness threshold (H=254 for overexposure detection) cannot distinguish between genuine sensor saturation and naturally high-reflectance surfaces. In subsequent research, we will attempt to combine semantic features with brightness features to identify better and delineate areas of brightness-missing regions.

C. Discussion on Generating Details

It is worth noting that diffusion models, by their nature, may generate plausible but non-original textures in regions

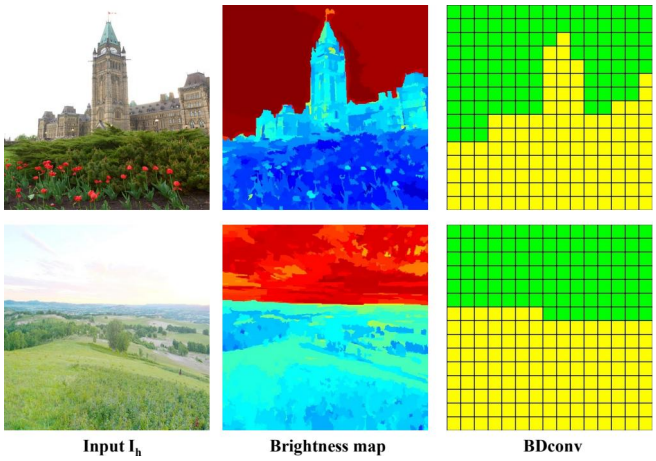


Fig. 10. Visualization of the brightness map B of the image and Brightness-Aware Dynamic Convolution. MPD effectively grades the luminance to apply different dynamic convolution kernels.

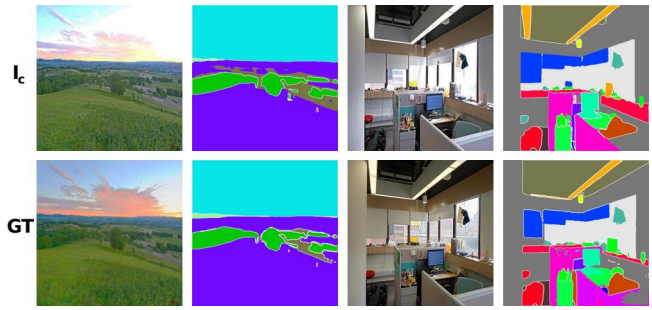


Fig. 11. Illustration of the corrected image I_c and the ground truth, and their corresponding semantic maps. Although the brightness-missing regions in the corrected image I_c have lost detailed information, semantic features can still be obtained and are consistent with the ground truth.

where pixel-level information is entirely missing, such as in extremely overexposed or underexposed regions. This is an inherent limitation of generative models, as they rely on learned data distributions and semantic priors to infer missing content. For example, in overexposed skies, the model might generate cloud-like patterns, or in underexposed vegetation, it might produce foliage details that align with semantic expectations but may not match the original scene. While this phenomenon is a known challenge in generative approaches, our method effectively addresses the core issue of restoring brightness-missing regions by integrating structural semantics, and exposure semantics. These priors guide the model to recover realistic and contextually consistent details, improving the quality of restored images compared to traditional methods. Our framework thus provides a robust solution to the problem of extreme exposure restoration, even within the inherent limitations of diffusion models. Future work could explore additional constraints or hybrid approaches to further refine the balance between plausibility and accuracy.

VI. Conclusion

This work proposes MPD, a novel framework for restoring images afflicted by extreme underexposure or overexposure.

(a) Input (b) Failure case

Fig. 12. Failure case. The method misjudges the white exterior walls that should exist in the building as overexposed regions (brightness-missing regions), resulting in the generation of repair content that does not match the actual texture on the normal white surface.

Diverging from conventional exposure correction techniques that predominantly adjust dynamic range, our method tackles the pivotal issue of brightness-missing regions resulting from grayscale truncation under extreme conditions. We introduce a novel multi-level priors mechanism that delivers comprehensive guidance: Low-level brightness priors facilitate pixel-adaptive correction via brightness-aware dynamic convolution, adeptly redistributing illumination while maintaining spatial consistency; High-level structural semantics priors preserve contextual coherence between restored and original content through cross-scale semantic attention; High-level exposure semantics priors establish cross-modal alignment between textual exposure prompts and visual features, thereby mitigating local over/under-correction artifacts. The synergistic integration of these complementary priors offers a fundamental solution to the joint optimization challenge of achieving semantic accuracy and exposure naturalness in extreme exposure restoration.

References

[1] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1780–1789.

[2] D. Liang, L. Li, M. Wei, S. Yang, L. Zhang, W. Yang, Y. Du, and H. Zhou, "Semantically contrastive learning for low-light image enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 1555–1563.

[3] D. Liang, Z. Xu, L. Li, M. Wei, and S. Chen, "Pie: Physics-inspired low-light enhancement," *International Journal of Computer Vision*, pp. 1–22, 2024.

[4] J. Huang, Y. Liu, F. Zhao, K. Yan, J. Zhang, Y. Huang, M. Zhou, and Z. Xiong, "Deep fourier-based exposure correction network with spatial-frequency interaction," in *European Conference on Computer Vision*. Springer, 2022, pp. 163–180.

[5] H. Wang, K. Xu, and R. W. Lau, "Local color distributions prior for image enhancement," in *European Conference on Computer Vision*. Springer, 2022, pp. 343–359.

[6] M. Afifi, K. G. Derpanis, B. Ommer, and M. S. Brown, "Learning multi-scale photo exposure correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9157–9167.

[7] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 177–1, 2017.

[8] S. Lee, G. H. An, and S.-J. Kang, "Deep recursive hdri: Inverse tone mapping using generative adversarial networks," in *proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 596–611.

[9] J. Kim, S. Lee, and S.-J. Kang, "End-to-end differentiable learning to hdr image synthesis for multi-exposure images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 1780–1788.

[10] Y. Niu, J. Wu, W. Liu, W. Guo, and R. W. Lau, "Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions," *IEEE Transactions on Image Processing*, vol. 30, pp. 3885–3896, 2021.

[11] Q. Yan, T. Hu, Y. Sun, H. Tang, Y. Zhu, W. Dong, L. Van Gool, and Y. Zhang, "Towards high-quality hdr deghosting with conditional diffusion models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

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

[13] 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.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[15] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *ICCV*, 2023.

[16] J. Bai, Y. Yin, Q. He, Y. Li, and X. Zhang, "Retinexmamba: Retinex-based mamba for low-light image enhancement," in *International Conference on Neural Information Processing*. Springer, 2024, pp. 427–442.

[17] Z. Li, F. Zhang, M. Cao, J. Zhang, Y. Shao, Y. Wang, and N. Sang, "Real-time exposure correction via collaborative transformations and adaptive sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2984–2994.

[18] Y. Li, K. Xu, G. P. Hancke, and R. W. Lau, "Color shift estimation-and-correction for image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[19] Y. Zhou, D. Liang, S. Chen, and S.-J. Huang, "Image lens flare removal using adversarial curve learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[20] Y. Guo, Y. Gao, B. Hu, X. Qian, and D. Liang, "Cmid: crossmodal image denoising via pixel-wise deep reinforcement learning," *Sensors*, vol. 24, no. 1, p. 42, 2023.

[21] 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.

[22] G. Eilertsen, R. K. Mantiuk, and J. Unger, "Single-frame regularization for temporally stable cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 176–11 185.

[23] Y. Li, D. Liang, T. Ding, and S.-J. Huang, "Structsr: Refuse spurious details in real-world image super-resolution," *arXiv preprint arXiv:2501.05777*, 2025.

[24] X. Lin, J. He, Z. Chen, Z. Lyu, B. Dai, F. Yu, Y. Qiao, W. Ouyang, and C. Dong, "Diffbir: Toward blind image restoration with generative diffusion prior," in *European conference on computer vision*. Springer, 2024, pp. 430–448.

[25] J. Zhou, T. Ding, T. Chen, J. Jiang, I. Zharkov, Z. Zhu, and L. Liang, "Dream: Diffusion rectification and estimation-adaptive models," *arXiv preprint arXiv:2312.00210*, 2023.

[26] 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.

[27] S. Yang, X. Chen, and J. Liao, "Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3190–3199.

[28] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," in *The Eleventh International Conference on Learning Representations*, 2023.

[29] Z. Wu, K. Li, H. Fan, Y. Yang, and C. ReLER, "Drafting and revision: advancing high-fidelity video inpainting," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 2063–2071.

[30] Z. Wu, K. Chen, K. Li, H. Fan, and Y. Yang, "Bvinet: Unlocking blind video inpainting with zero annotations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 14 017–14 027.

[31] Z. Wu, H. Xuan, C. Sun, W. Guan, K. Zhang, and Y. Yan, "Semi-supervised video inpainting with cycle consistency constraints," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 586–22 595.

[32] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[33] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS

[34] W.-K. Ching and M. K. Ng, "Markov chains," *Models, algorithms and applications*, 2006.

[35] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=0xiJLKH-ufZ

[36] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

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

[38] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.

[39] 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.

[40] L. Yuan and J. Sun, "Automatic exposure correction of consumer photographs," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*. Springer, 2012, pp. 771–785.

[41] J. Chen, X. Wang, Z. Guo, X. Zhang, and J. Sun, "Dynamic region-aware convolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8064–8073.

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

[43] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *Journal of Visual Communication and Image Representation*, vol. 26, pp. 132–145, 2015.

[44] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[45] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[46] N. E. Nsampi, Z. Hu, and Q. Wang, "Learning exposure correction via consistency modeling." in *BMVC*, 2021, p. 12.

[47] K.-F. Yang, C. Cheng, S.-X. Zhao, H.-M. Yan, X.-S. Zhang, and Y.-J. Li, "Learning to adapt to light," *International Journal of Computer Vision*, vol. 131, no. 4, pp. 1022–1041, 2023.

[48] R. Cui and L. Zhang, "Unice: Training a universal image contrast enhancer," *arXiv preprint arXiv:2507.17157*, 2025.

[49] T. Wang, K. Zhang, Y. Zhang, W. Luo, B. Stenger, T. Lu, T.-K. Kim, and W. Liu, "Lldiffusion: Learning degradation representations in diffusion models for low-light image enhancement," *Pattern Recognition*, vol. 166, p. 111628, 2025.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.

[52] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

[53] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *CVPR 2011*. IEEE, 2011, pp. 97–104.

[54] A. Mittal, R. Soundararajan, and A. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, pp. 209–212, 2013.

[55] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, pp. 3474–3486, 2021.
[56] 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.
[57] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, 2023.

Biography Section

Xinyue Zhao received her M.S. degree in Mechanical Engineering from Zhejiang University, China in 2008, and her Ph.D degree in Graduate School of Information Science and Technology from Hokkaido University, Japan in 2012. She is currently an associate professor in the School of Mechanical Engineering, Zhejiang University, China. Her research interests include machine vision and image processing. She has published nearly 50 peer reviewed journal papers.

Andong Zhang received the BS degree in computer science and technology from the Hangzhou Dianzi University, China, in 2024. He is currently working toward the master's degree with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His research interests include low-level computer vision tasks such as image enhancement.

Zhengyan Xu received the BS degree in computer science and technology from the Hefei University of Technology, China, in 2021. He received the M.S. degree in Computer Science and Technology from Nanjing University of Aeronautics and Astronautics, China, in 2025. He is currently pursuing a Ph.D. in Computer Science and Technology at the School of Computer Science and Technology, Beijing Institute of Technology. His research interests include medical image segmentation, multimodal fusion, and computer vision.

Yachao Li received the BS degree in Software Engineering and the M.S. degree in Computer Science and Technology from Nanjing University of Aeronautics and Astronautics, China, in 2022 and 2025. His research interests include low-level image enhancement, such as super-resolution, colorization.

Zaixing He received his B.Sc. and M.Sc. degrees in Mechanical Engineering from Zhejiang University, China in 2006 and 2008, respectively. He received his Ph. D. degree in 2012 from the Graduate School of Information Science and Technology, Hokkaido University, Japan. He is currently an associate professor in the School of Mechanical Engineering, Zhejiang University. His research interests include robotic/machine vision, Visual intelligence of manufacturing equipment, and optical-based measurement. He has published over 50 peer reviewed papers in prestigious journals such as IEEE TRO, IEEE/ASME TMech, IEEE TIE, TII, TIM, Pattern Recognition, Optics Letters, Neurocumputing etc. He served as Lead Guest Editor or Guest Editor of several journals including IEEE TCE and Mathematics, and Program Chair or TPC of several international conferences. He is a senior member of IEEE.

Dong Liang received a BS degree in telecommunication engineering and an MS degree in circuits and systems from Lanzhou University, China, in 2008 and 2011. He received PhD from Hokkaido University, Japan, in 2015. He is a professor at the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA). His research interests include machine learning and computer vision. He has published several research papers in IEEE TPAMI/TIP/TNNLS/TMM/TGRS, International Journal of Computer Vision, Pattern Recognition, and ICCV/ECCV/AAAI/IJCAI/ISMAR/ICME. He is an Associate Editor of The Visual Computer, and a senior member of IEEE.