GMODiff: One-Step Gain Map Refinement with Diffusion Priors for Efficient HDR Reconstruction

Tao Hu[1], Weiyu Zhou[1], Yanjie Tu[1], Wei Dong[3], Peng Wu[1], Qingsen Yan[1,2] *, and Yanning Zhang[1]

[1] Northwestern Polytechnical University, Xi'an 710072, China
[2] The Shenzhen Research Institute of Northwestern Polytechnical University, Shenzhen 518057, China
[3] Xi'an University of Architecture and Technology, Xi'an 710055, China
taohu@mail.nwpu.edu.cn, qingsenyan@nwpu.edu.cn

Abstract. Pre-trained Latent Diffusion Models (LDMs) have recently shown strong perceptual priors for low-level vision tasks, making them a promising direction for multi-exposure High Dynamic Range (HDR) reconstruction. However, directly applying LDMs to HDR remains challenging due to: (1) limited dynamic-range representation caused by 8-bit latent compression, (2) high inference cost from multi-step denoising, and (3) content hallucination inherent to generative nature. To address these challenges, we introduce GMODiff, a **g**ain **m**ap-driven **o**ne-step **diff**usion framework for multi-exposure HDR reconstruction. Instead of reconstructing full HDR content, we reformulate HDR reconstruction as a degradation-aware Gain Map (GM) refinement problem, where the GM encodes the extended dynamic range while retaining the same bit depth as LDR images. We initialize the denoising process from an informative regression-based estimate rather than pure noise, allowing the model to generate high-quality GMs in a single denoising step. Furthermore, recognizing that regression-based models excel in content fidelity while LDMs favor perceptual quality, we leverage regression priors to guide both the denoising process and latent decoding of the LDM, suppressing hallucinations while preserving structural accuracy. Extensive experiments demonstrate that our GMODiff performs favorably against several state-of-the-art methods and is 100× faster than previous LDM-based methods.

Keywords: HDR reconstruction · Diffusion model · Gain map

1 Introduction

Multi-exposure High Dynamic Range (HDR) imaging fuses Low Dynamic Range (LDR) images of varying exposures to recover scene radiance, but suffers from motion-induced misalignment in dynamic scenes. Recent Deep Neural Networks (DNNs), including Convolutional Neural Networks (CNNs) [43,46,50] and Vision

* Corresponding author.

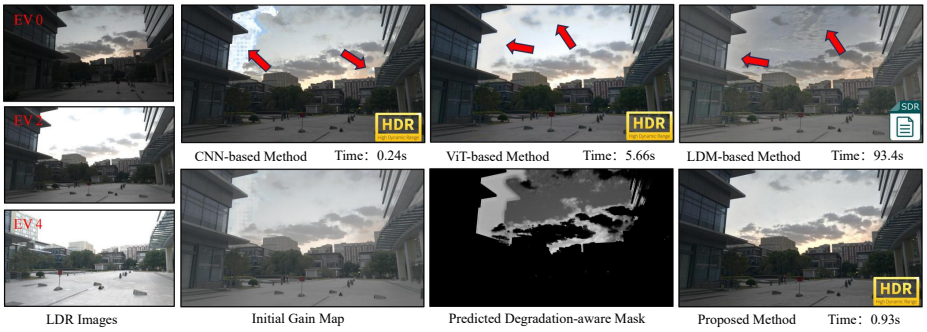

Fig. 1: Compared with CNN-based [46] and ViT-based [38] methods, the LDM-based method UltraFusion [8] generates perceptually compelling HDR-like images, highlighting the potential of LDMs for HDR reconstruction. However, it incurs high computational cost and may introduce hallucination artifacts that compromise physical fidelity.

Transformers (ViTs) [28, 38, 45], have greatly improved HDR reconstruction in dynamic scenes. Despite their advancements, these methods simply learn a mapping network between LDR-HDR paired data using pixel-wise losses [41], which often yields perceptually unsatisfactory results. This limitation is particularly pronounced when essential content is missing in severely overexposed regions due to object or camera motion.

Recent advances in Latent Diffusion models (LDMs) [9, 15], particularly large-scale pre-trained text-to-image (T2I) models [34, 35] trained on billions of image-text pairs, possess powerful natural image priors. These priors make LDMs a promising solution for addressing the perceptual shortcomings of current HDR reconstruction methods, especially in mitigating motion-induced ghosting and other complex artifacts. However, directly applying LDMs to high-quality HDR image reconstruction remains non-trivial due to three key challenges: **(1) Limited Dynamic Range Representation:** The autoencoders of pre-trained LDMs compress linear HDR distributions into an 8-bit perceptual latent space, limiting their capacity to represent the dynamic range of HDR content [13, 40]. **(2) High Inference Cost:** LDMs demand a substantial number of iteration steps in the denoising process, which is time-consuming even with a high-end GPU card [20,47]. **(3) Content Hallucination:** The generative nature of LDMs can lead to hallucinated content that deviates from realistic scenes. As shown in Fig. 1, UltraFusion [8] produces human-perceptually consistent HDR-like content with a compressed dynamic range but suffer from high latency and hallucination.

These limitations motivate a paradigm shift in leveraging LDMs for HDR reconstruction. Our work draws from three key insights: **First,** inspired by the observation that an HDR image can be decomposed into a LDR component and a 8 bit-depth Gain Map (GM) [1,2,12] that encodes the extended dynamic range, we reframe HDR reconstruction as a conditionally guided GM refinement task to fully exploit pre-trained LDM priors without domain mismatches. **Second,** since LDR inputs already provide rich structural and semantic cues, we initialize

the denoising process from an informative prior (*e.g.*, a regression-based GM estimate) rather than pure noise, significantly accelerating inference. **Third,** recognizing that regression-based models excel in content fidelity while LDMs favor perceptual quality, we leverage regression-based priors to guide the LDM's denoising and latent decoding processes.

Building on these principles, we propose GMODiff, a gain map-driven one-step diffusion model for multi-exposure HDR reconstruction. To the best of our knowledge, GMODiff is the first LDM-based method designed for multi-exposure HDR reconstruction. Rather than training a LDM to directly reconstruct full HDR radiance, we adopt a two-stage fine-tuning strategy to adapt a pre-trained multi-step LDM for single-step GM refinement. Specifically, in the first stage, we train a Degradation-aware Regressor Network (DaReg) via a dual-learning strategy to produce an initial GM and extract spatial embeddings that highlight regions where the estimate is unreliable. This learning scheme provides the LDM with an informative prior for initialization and directs its attention to ambiguous or error-prone areas, significantly accelerating convergence and improving structural fidelity. In the second stage, we fine-tune the LDM to perform one-step denoising conditioned on these regression-based priors, enabling accurate GM correction and high-fidelity HDR reconstruction at dramatically reduced computational cost. In summary, our main contributions are threefold:

- We propose GMODiff, a one-step diffusion framework that reframes HDR reconstruction as a conditionally guided gain map refinement task, fully leveraging pre-trained LDMs without redesigning the latent space.
- We design a degradation-aware refinement pipeline that seamlessly integrates regression-based fidelity with diffusion-driven perceptual enhancement: the DaReg produces an initial gain map and degradation-aware embeddings, which guide both the denoising process and the latent decoding of the LDM, effectively suppressing hallucinations while preserving structural accuracy.
- Extensive experiments demonstrate that the proposed GMODiff delivers superior visual quality with significantly reduced ghosting and artifacts, while achieving high inference efficiency.

2 Related Work

2.1 Multi-exposure HDR Reconstruction

Multi-exposure HDR imaging has advanced significantly, with methods broadly divided into traditional and deep learning-based categories. Traditional methods address motion artifacts via motion rejection [11], registration [4, 22], and patch matching [17, 36], but their effectiveness hinges on preprocessing accuracy, often failing with large/complex motion. In contrast, state-of-the-art HDR reconstruction methods now predominantly leverage the powerful nonlinear modeling capabilities of DNNs. Recent advances incorporate attention mechanisms [19, 45, 46] and ViT architectures [28, 38, 45, 51] to improve feature correspondence and capture long-range dependencies. Meanwhile, optical flow–guided approaches [21, 23]

explicitly address misalignment by warping input frames according to estimated motion fields. Moreover, generative models, including Generative Adversarial Networks (GANs) [31] and diffusion models [8, 20, 40, 47, 49], have been successfully adapted to synthesize photorealistic HDR content, even from imperfect or severely degraded inputs. Additional innovative frameworks, such as [18, 26, 30], have further expanded the scope of DNN applications in this field, introducing novel architectures and training strategies to improve reconstruction quality. Despite these advances, most existing methods either learn a direct LDR-to-HDR mapping using pixel-wise losses, or train generative models from scratch on limited datasets. These often yield unsatisfactory results, especially when critical content is missing in severely over/underexposed regions due to motion.

2.2 Gain Map-based HDR Reconstruction

To support consistent and device-adaptive HDR image rendering, several recent industry standards have introduced a dual-layer image representation [1, 2, 12]. This format jointly stores a standard LDR image together with an auxiliary gain map, enabling flexible reconstruction of HDR content across diverse display environments. Recently, several GM-based HDR reconstruction methods have emerged. Liao *et al*. [27] introduced the Gain Map–based Inverse Tone Mapping (GM-ITM) task, shifting the learning objective from direct HDR prediction to gain map estimation, thereby achieving greater flexibility. Canham *et al*. [6] proposed an alternative to the multiplicative GM by introducing a pixel-wise exponent map, and in their subsequent work, they further introduced an MLP-based GM encoder [5], which demonstrated improved efficiency in HDR recovery. More recently, Guan *et al*. [13] leveraged two cascaded diffusion models to separately synthesize the LDR image and its corresponding GM for HDR image generation. Compared to traditional HDR reconstruction methods, the GM exhibits a more balanced distribution [27] and offers greater flexibility during the reconstruction process. Motivated by these benefits, we reformulated the multi-exposure HDR reconstruction task as a conditionally guided GM estimation problem, which allows us to fully exploit pre-trained LDM priors without suffering from domain mismatches.

2.3 Diffusion Models

Diffusion models [34, 35] have emerged as a powerful class of generative frameworks, capable of synthesizing high-fidelity data through iterative denoising of random noise. They have already demonstrated impressive performance across a wide range of low-level vision tasks-such as super-resolution [42], denoising [10], dehazing [24, 25], JPEG artifact removal [14], image enhancement [48], and deblurring [44], thanks to their ability to model fine-grained details and generate visually compelling outputs, often surpassing traditional image restoration methods. Despite their success, the practical deployment of diffusion models remains hindered by their high computational cost. To mitigate this bottleneck, recent efforts [29, 37] have focused on accelerating diffusion by reducing the number of

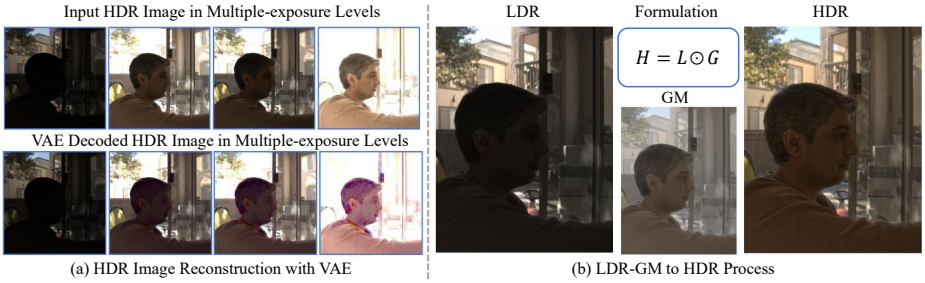

Fig. 2: (a) Limitations of the vanilla LDMs in generating HDR content. The limitation of the LDM VAE in encoding and decoding an HDR image, visualized in multiple exposure levels, which reveals a significant fidelity loss, especially in the shadow. (b) The GM encodes pixel-wise dynamic range adjustments for the LDR image and can be used to reconstruct the HDR image via element-wise multiplication.

denoising steps. However, aggressive step reduction typically leads to a noticeable degradation in output quality. Recently, img2img [32] proposed a general framework that adapts pre-trained single-step diffusion models to new tasks and domains via adversarial learning. Meanwhile, several recent studies [10, 14, 42] have explored single-step diffusion models for low-level vision tasks. These methods enable efficient inference while retaining the rich generative priors embedded in large-scale pre-trained diffusion models. Nevertheless, directly applying this efficient diffusion paradigm to HDR generation remains challenging. Most pre-trained diffusion models rely on autoencoders trained on 8-bit LDR images, which compress HDR's high bit-depth into a perceptually encoded, low-dimensional latent space, severely limiting their ability to capture the full dynamic range of the real world.

3 Methodology

As shown in Fig. 3, GMODiff is trained in two stages. During the first stage (Sec. 3.2), we pretrain a Degradation-aware Regressor Network (DaReg) to predict an initial gain map and degradation-aware embeddings from the LDR inputs. During the second stage (Sec. 3.3), we fine-tune a pre-trained LDM using these regression priors to refine the initial gain map and reconstruct a high-quality HDR image.

3.1 Preliminaries

As shown in Fig. 2 (a), while the variational autoencoder (VAE) of a LDM can encode visually plausible LDR representations that resemble HDR content under standard viewing conditions, it fundamentally fails to preserve true high dynamic range-particularly in highlights and shadows. Since the denoising network operates in an 8-bit perceptual latent space, a naive solution to enable

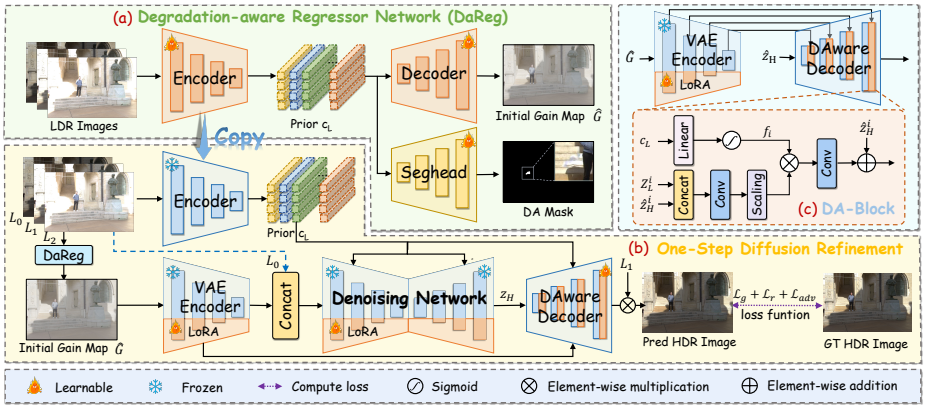

Fig. 3: Overview of the proposed GMODiff framework. (a) We train a DaReg via a dual-learning strategy to produce two regression-based priors: an initial gain map $\hat{G}$ and spatial embeddings c_L. The c_L implicitly encode regions where the $\hat{G}$ is unreliable, serving as a degradation-aware guidance prior. (b) One-Step Diffusion Refinement initializes the denoising process from $\hat{G}$ and performs single-step denoising conditioned on c_L to generate a high-fidelity GM latent code Z_H. (c) The DA Decoder leverages encoder features from the initial gain map $\hat{G}$, guided by c_L, to recover fine image details while avoiding the introduction of artifacts inherent in $\hat{G}$.

HDR generation would be to fine-tune both the VAE and the denoising network on HDR data. However, this approach demands large-scale HDR training sets and discards the rich generative priors learned from massive LDR images [40].

Recent advances in display technology have popularized dual-layer HDR formats [2, 12], which decompose an HDR image $\mathbf{H}$ into an LDR base layer $\mathbf{L}$ and an 8-bit gain map $\mathbf{G}$. The gain map is obtained by pixel-wise division of $\mathbf{H}$ by $\mathbf{L}$, followed by logarithmic compression. This representation has also shown promise for diffusion-based HDR generation. Specifically, Guan *et al.* [13] employed two cascaded diffusion models to separately synthesize the LDR base image and its corresponding gain map for HDR image generation. As illustrated in Fig. 2 (b), the HDR image can be recovered from the dual-layer representation as:

$$\mathbf{H} = (\mathbf{L} + \alpha) \cdot \exp_2 \left(1 + \mathbf{G} \cdot Q_{\max}\right), \tag{1}$$

where $\mathbf{G} \in [0, 1]$ is the normalized gain map, $\exp_2(\cdot)$ denotes the inverse log transform, $Q_{\max}$ corresponds to the maximum gain value, and $\alpha = 1/64$ is a scaling constant that preserves detail in near-black regions. Crucially, both the LDR image and gain map are represented at 8-bit depth, allowing us to directly exploit pre-trained LDMs without VAE retraining or latent-space redesign.

3.2 Degradation-aware Regressor Network

The first stage produces two regression-based priors: an initial gain map $\hat{\mathbf{G}}$ and multi-scale spatial embeddings $\mathbf{c}_L$. These priors are used to initialize and con-

dition the one-step diffusion refinement, ensuring high structural fidelity while enabling perceptual enhancement.

Overall Architecture. Our DaReg adopts a NAFNet backbone [7] augmented with a lightweight degradation-aware segmentation head. Given the input LDR sequence $\{L_1, L_2, \ldots, L_N\}$, we first apply gamma correction to each L_i to obtain its HDR version H_i, and form a six-channel tensor $X_i = [L_i, H_i]$ by channel-wise concatenation. Following [28, 46], an implicit alignment module processes the LDR inputs and fuses multi-exposure features before feeding them into the network. As shown in Fig. 3, DaReg encodes the fused features into a latent representation $\mathbf{c}_L$. For diffusion-based refinement, the spatial feature map is flattened and projected into a sequence of spatial embeddings:

$$\mathbf{c}_L = \left\{ \mathbf{c}_{L_k} \in \mathbb{R}^d \right\}_{k=1}^{K}, \tag{2}$$

where K denotes the number of spatial locations, and each $\mathbf{c}_{L_k}$ corresponds to the feature representation at one spatial location. For notational simplicity, we use $\mathbf{c}_L$ to denote this latent representation in both its spatial form for initial gain-map reconstruction and its tokenized form for subsequent diffusion refinement. The decoder reconstructs the initial gain map using the spatial form of $\mathbf{c}_L$, while the segmentation head predicts a pixel-wise reliability map for degraded regions (*e.g.*, due to motion or saturation). After the first stage of training, the predicted reliability map is fused with the tokenized $\mathbf{c}_L$ to provide explicit conditional cues for uncertain regions during subsequent refinement.

Initial GM Reconstruction. Given an input LDR sequence, the encoder extracts hierarchical features, where each stage consists of multiple Nonlinear Activation-Free Blocks [7]. A symmetric decoder then reconstructs the initial gain map $\hat{\mathbf{G}}$ from the spatial form of $\mathbf{c}_L$ as $\hat{\mathbf{G}} = D\left(\mathbf{c}_L; \theta_D\right)$, where each decoder stage receives skip-connected features from the corresponding encoder stage to progressively reconstruct the initial GM.

During training, we compute the ground-truth gain map $\mathbf{G}$ using Eq. (1), where $\mathbf{L}$ is the reference frame of the LDR sequence. To enable direct supervision with HDR images and ensure differentiability at zero, we replace the logarithmic compression with the μ-law function. Accordingly, the $\exp_2(\cdot)$ in Eq. (1) is substituted by the inverse μ-law mapping: $R(x) = \frac{e^{x \cdot \log(1+\mu)} - 1}{\mu}$, with μ=100. We supervise the initial gain map prediction $\hat{\mathbf{G}}$ via an $\mathcal{L}_1$ regression loss:

$$\mathcal{L}_{\text{gm}} = \|\hat{\mathbf{G}} - \mathbf{G}\|_1. \tag{3}$$

Additionally, we enforce consistency between the reconstructed HDR image $\hat{\mathbf{H}}$ (from $\hat{\mathbf{G}}$) and the ground-truth $\mathbf{H}$:

$$\mathcal{L}_r = \left\| \mathcal{T}(\mathbf{H}) - \mathcal{T}(\hat{\mathbf{H}}) \right\|_1 + \lambda \left\| \phi_{i,j}(\mathcal{T}(\mathbf{H})) - \phi_{i,j}(\mathcal{T}(\hat{\mathbf{H}})) \right\|_1, \tag{4}$$

where $\mathcal{T}(\cdot)$ denotes the μ-law function, $\phi_{i,j}(\cdot)$ is the j-th feature map of VGG19 after the i-th max-pooling layer, and $\lambda = 1\text{e}{-2}$ is a weighting hyperparameter.

Degradation-Aware Embedding. By incorporating a reconstruction objective, DaReg implicitly learns regression-based priors that facilitate dynamic range expansion. To better guide the LDM in refining the initial GM, we additionally introduce a segmentation head to predict a degradation-aware mask, which allows the refinement process to focus on uncertain regions and avoids hallucinating implausible details in well-reconstructed areas.

Specifically, given an input LDR sequence, we extract multi-scale latent representations with the DaReg encoder as before. For the degradation-aware mask M, we process the encoder features to the same number of channels using 1×1 convolutions, and upsample them to the original resolution for concatenation to produce $\bar{\mathbf{c}}_L$. The segmentation head $h(\cdot;\phi)$ processes $\bar{\mathbf{c}}_L$ through convolutional layers with 64 hidden channels, followed by a sigmoid activation, and outputs a predict mask: $\hat{M} = h(\bar{\mathbf{c}}_L;\phi)$, where $\hat{M} \in [0,1]$ denotes the predicted degradation probability at each pixel p. Notably, to prevent the gradient of the segmentation head from affecting the reconstruction accuracy, we perform a detach operation before feature processing. The obtained degradation-aware mask is then compressed through a lightweight CNN network in the next training stage and fused with the original $\mathbf{c}_L$ to serve as the conditional guidance for the LDMs.

We obtain the ground-truth mask M using a simple yet effective pixel-consistency criterion:

$$M(p) = \begin{cases} 1 & \text{if } \left\| \mathcal{T}(\mathbf{H}(p)) - \mathcal{T}(\hat{\mathbf{H}}(p)) \right\|_1 > \sigma, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where p denotes a pixel location, $\|\cdot\|_1$ averages the absolute difference across RGB channels to yield a scalar, and $\sigma = 4/255$ is a fixed threshold. This formulation is motivated by the observation that static regions yield consistent reconstructions-even with simple image blending-whereas motion or overexposure causes significant pixel-level discrepancies. The loss function adopted is the binary cross-entropy (BCE) loss between $\hat{M}$ and M.

3.3 Prior-Guided One-Step Diffusion Refinement

In the second stage, our goal is to efficiently leverage the strong image priors of pre-trained LDMs to address degradations in the initial GM (*e.g.*, ghosting and artifacts), thereby reconstructing high-quality HDR images.

One-step Denoising Process. Based on the prior analysis, we formulate GM refinement as a conditionally guided image-to-image denoising process, which enables the adoption of an efficient one-step denoising strategy [42] and accelerates convergence. Our GMODiff first encodes the initial GM into the latent space via an encoder E_θ, resulting in $\mathbf{Z}_L = E_\theta(\hat{\mathbf{G}})$. A single-step denoising operation $F(\mathbf{Z}_L;\mathbf{c})$ is then performed to predict the noise $\hat{\epsilon}$, which enables us to derive the estimated high-quality latent representation $\hat{\mathbf{Z}}_H$:

$$\hat{\mathbf{Z}}_H = F(\mathbf{Z}_L;\mathbf{c}) \triangleq \frac{\mathbf{Z}_L - \sqrt{1-\bar{\alpha}_{T_L}}\varepsilon_\theta\left(\mathbf{Z}_L;\mathbf{c},T\right)}{\sqrt{\bar{\alpha}_T}}, \tag{6}$$

where ε_θ denotes the denoising network, c is the guidance condition, $T \in [0, T]$ is the predefined diffusion timestep, and $\bar{\alpha}$ represents the predefined noise schedule. Specifically, we append trainable LoRA [16] layers to the encoder E_ϕ and the diffusion network ϵ_ϕ, fine-tuning them into E_θ and ϵ_θ using our training data. The condition $\mathbf{c}$ includes the regression prior $\mathbf{c}_L$ and the underexposed LDR image L_0. $\mathbf{c}_L$ is injected into the hidden layers of the denoising network, whereas L_0 is first encoded into a feature map, then concatenated with $\mathbf{Z}_L$ and fed into the denoising network, with the number of channels in the first convolutional layer adjusted accordingly.

Degradation-aware Decoder. The pre-trained LDM performs denoising in a highly compressed latent space (*e.g.*,$\mathcal{R} \in \mathbb{R}^{4\times\frac{H}{8}\times\frac{W}{8}}$), which can result in detail distortion in generated content [32]. To address this issue, we propose a Degradation-aware Decoder that leverages the encoder features of $\hat{\mathbf{G}}$ under the guidance of degradation-aware priors to facilitate refinement of detail while suppressing artifacts inherent in the initial GM $\hat{\mathbf{G}}$.

Specifically, we obtain multi-scale features $\mathbf{Z}_L^i$ of the initial GM $\hat{\mathbf{G}}$ via the VAE encoder through forward propagation, while $\hat{\mathbf{Z}}_H$ is derived from a one-step diffusion process. Although the initial gain map, reconstructed via a regression approach, endows $\mathbf{Z}_L^i$ with rich detail fidelity, it may introduce noticeable artifacts. To address this, we propose a Degradation-aware Refinement Block guided by the embedding $\mathbf{c}_L$, which leverages degradation-aware embeddings to identify unreliable regions in $\mathbf{Z}_L^i$. As illustrated in Fig. 3 (c), the Degradation-aware Refinement Block first concatenates the decoder intermediate features $\hat{\mathbf{Z}}_H^i$ with $\mathbf{Z}_L^i$ and applies a 3×3 convolution to obtain refined features $\hat{\mathbf{Z}}^i$:

$$\hat{\mathbf{Z}}^i = \mathrm{Conv}_{3\times3}\left([\hat{\mathbf{Z}}_H^i, \mathbf{Z}_L^i]\right), \tag{7}$$

where $[\cdot]$ denotes channel-wise concatenation. The $\mathbf{c}_L$ is then passed through an MLP followed by a sigmoid activation to produce a channel-wise modulation vector:

$$\begin{aligned} \mathbf{f}_i &= \mathrm{Sigmoid}\left(\mathrm{MLP}(\mathbf{c}_L)\right), \\ \hat{\mathbf{Z}}_H^{i+1} &= \mathrm{Conv}_{3\times3}\left(\hat{\mathbf{Z}}^i \odot \mathbf{f}_i\right) + \hat{\mathbf{Z}}_H^i, \end{aligned} \tag{8}$$

where $\odot$ denotes channel-wise multiplication, and $\hat{\mathbf{Z}}_H^{i+1}$ represents the updated decoder feature after detail modulation and residual connection.

Training Strategy During the training phase, the loss function comprises the GM reconstruction loss and HDR reconstruction loss. Following [14], we also compute the adversarial loss $\mathcal{L}_{\mathrm{adv}}$ between the predicted latent gain map $\hat{\mathbf{Z}}_H$ and the GT latent gain map $\mathbf{Z}_H$, where the discriminator consists of a frozen pre-trained denoising UNet encoder and two trainable MLP layers. The total objective for decoder training is:

$$\mathcal{L}_{diff} = \mathcal{L}_{gm} + \mathcal{L}_r + \lambda\mathcal{L}_{adv}, \tag{9}$$

where λ denotes the balancing parameter, which is empirically set to 0.001.

Additionally, considering that the initial GM $\hat{\mathbf{G}}$ obtained in the first stage may overfit to the training set, which impairs GMODiff's ability to handle its

Table 1: Quantitative comparison on the in-distribution test split of our collected dataset. Best and second-best results per metric are highlighted in red and blue, respectively.

Metric	DeepHDR [43]	AHDRNet [46]	SAFNet [31]	HDR-Trans [28]	SCTNet [38]	AFUNet [26]	LE-Diff [40]	GMDD [13]	DiffHDR [47]	Ours
PU21-PSNR↑	32.91	35.37	36.53	36.30	35.47	36.91	-	16.31	35.43	36.68
PU21-SSIM↑	0.9814	0.9861	0.9876	0.9868	0.9871	0.9889	-	0.7740	0.9861	0.9882
HDR-VDP2↑	67.98	68.74	68.84	68.16	68.92	70.77	-	63.01	68.79	70.77
DISTS↓	0.0221	0.0172	0.0112	0.0140	0.0121	0.0120	-	0.0856	0.0125	0.0103
MANIQA↑	0.5786	0.5828	0.5817	0.5804	0.5857	0.5820	0.5621	0.5738	0.5834	0.5881
MUSIQ↑	59.59	59.77	60.59	60.14	60.59	60.66	57.07	57.88	60.39	60.96
CLIPIQA↑	0.3946	0.4085	0.4199	0.4105	0.4183	0.4164	0.3675	0.4225	0.4093	0.4253

Table 2: Quantitative comparisons on the Prabhakar *et al.*'s [33] dataset. Best and second-best results per metric are highlighted in red and blue, respectively.

Metric	DeepHDR [43]	AHDRNet [46]	SAFNet [31]	HDR-Trans [28]	SCTNet [38]	AFUNet [26]	LE-Diff [40]	GMDD [13]	DiffHDR [47]	Ours
PU21-PSNR↑	23.06	26.30	26.13	25.18	27.41	27.78	-	15.03	25.31	26.86
PU21-SSIM↑	0.8775	0.9413	0.9410	0.9401	0.9454	0.9398	-	0.7577	0.9389	0.9419
HDR-VDP2↑	63.79	67.67	66.59	64.95	64.78	66.46	-	63.68	66.78	67.91
DISTS↓	0.0810	0.0433	0.0521	0.0644	0.0407	0.0589	-	0.1383	0.0395	0.0379
MANIQA↑	0.5222	0.5450	0.5410	0.5319	0.5412	0.5305	0.4413	0.5011	0.5369	0.5469
MUSIQ↑	52.95	52.55	53.25	54.61	53.59	52.70	34.57	42.41	53.67	55.48
CLIPIQA↑	0.3938	0.4387	0.4478	0.4489	0.4513	0.4395	0.3767	0.3861	0.4474	0.4602

degradations, we propose a group training strategy for GMODiff to learn robust degradation mitigation. Specifically, we divide the training set into N groups and train $N+1$ DaRegs accordingly, with the extra model trained on the entire training dataset. During GMODiff training, we randomly select one DaReg to provide $\hat{\mathbf{G}}$ and $\mathbf{c}_L$. The training objective for GMODiff is ultimately defined as the expected loss over such a random selections:

$$\mathcal{L} = \mathbb{E}_k \sim \text{Uniform}(0, N) \left[\mathcal{L}_{diff} \left(\hat{G}_k, \mathbf{c}_{L,k}; H; G \right) \right], \tag{10}$$

where k is the index of a randomly selection.

4 Experiments

4.1 Experimental Settings

Datasets. Scenes and motion variations within a single dataset are limited. To better validate the model's generalization ability under diverse real-world conditions, we collect publicly available real-world datasets [21,23,38] commonly used for multi-exposure HDR reconstruction, which cover a wide range of illumination and motion conditions. We follow the official training and test splits of all datasets. For evaluation, we conduct quantitative comparisons on datasets with ground truth, including the test split of our training data (in-distribution) and the Prabhakar *et al.* [33] dataset (out-of-distribution) for real-world robustness testing. In addition, we use the Sen *et al.* dataset [36] and the Tursun *et al.* dataset [39] solely for qualitative comparisons, as they do not provide ground-truth HDR images.

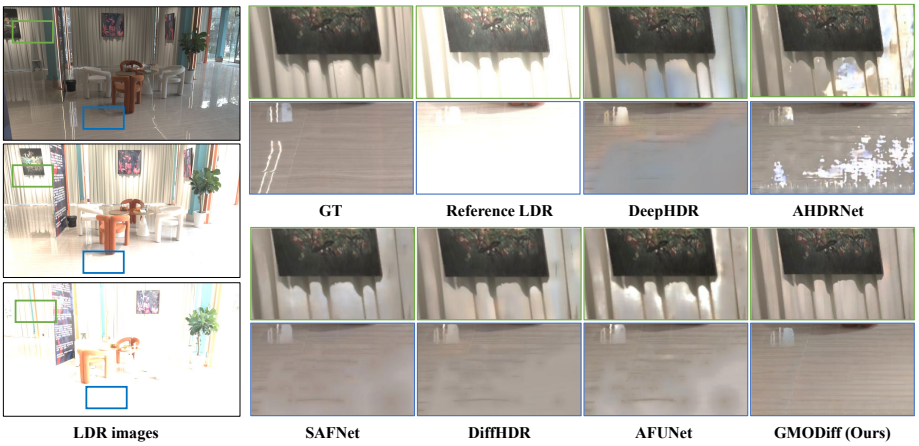

Fig. 4: Visual comparisons are conducted on the test set, with a focus on zoomed-in local regions of HDR images reconstructed by our method and competing techniques. Compared with existing methods, GMODiff more faithfully recovers local structures and illumination details while preserving visually natural brightness transitions.

Implementation Details. We implemented our model in PyTorch and used the `accelerate` library for mixed-precision training with FP16. In stage one, we optimize DaReg using the Adam optimizer with a learning rate of 2e-4. The training data is preprocessed by extracting 256×256 patches, and we use a batch size of 16. DaReg adopts the NAFNet architecture [7], with the encoder and decoder comprising [2, 4, 4, 8] and [2, 2, 2, 2] NAFBlocks per stage, respectively. The model is trained for 200K iterations on NVIDIA RTX 4090 GPUs, with a per-GPU batch size of 16. After training, DaReg is frozen, and the segmentation head is independently trained for 5K iterations. In Stage 2, we fine-tune Stable Diffusion 2.1 [34] using LoRA with a rank of 16, while keeping DaReg and the segmentation head frozen as fixed modules. We employ the AdamW optimizer with a learning rate of 5e-5, a weight decay of 1e-10, and a per-GPU batch size of 4. This stage is trained for 300K iterations on NVIDIA RTX 4090 GPUs.

Evaluation Metrics. To ensure fair and comprehensive comparison, we employ both perceptual and distortion-based image quality metrics. For full-reference perceptual evaluation, we use DISTS, which assess perceptual similarity via features from pre-trained deep networks. For no-reference assessment, we adopt MANIQA, MUSIQ, and CLIPIQA, all trained to predict human-perceived quality. In particular, as these metrics are not HDR-specific, we computed them in the tone-mapped domain using a standard μ-law operator. For distortion-based metrics, we include PU-PSNR and PU-SSIM, specifically tailored for HDR images, through the official implementation [3] with a peak luminance of 1000 nits. We compute HDR-VDP2 on linear HDR images with the "rgb-native" color encoding. The data is treated as display-referred, scaled to a peak luminance of 1000 nits, with a viewing angle of 30° and viewing distance of 0.55 m.

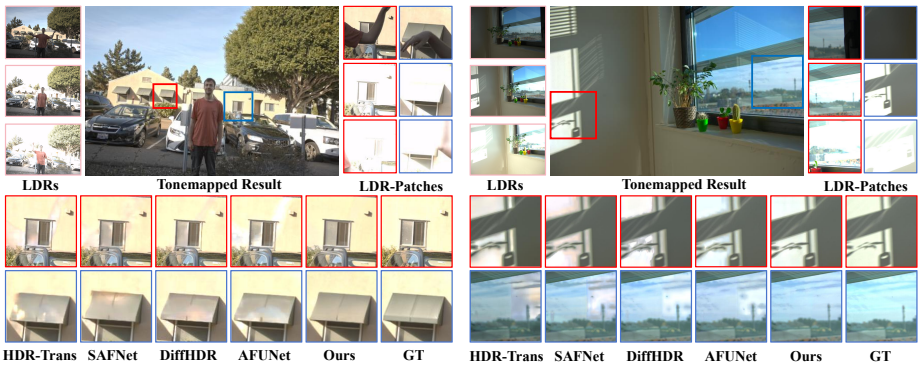

Fig. 5: Visual comparisons are conducted on testing data , focusing on zoomed-in local areas of the HDR images estimated by our method and the compared techniques. Our model demonstrates the ability to generate HDR images of superior quality.

Fig. 6: Qualitative comparison of HDR reconstruction results with state-of-the-art LDM-based methods. Our GMODiff outperforms existing LDM-based methods in recovering fine details, preserving physical illumination consistency and suppressing artifacts in overexposed/motion regions.

4.2 Comparison with SOTA Methods

Compared Methods. We compare GMODiff against state-of-the-art CNN- and Transformer-based methods specifically designed for multi-exposure HDR image reconstruction, including DeepHDR [43], AHDRNet [46], SCTNet [38], and HDR-Trans [28]. In addition, we also evaluate against recent advanced approaches: the diffusion-based DiffHDR [47], the optical-flow-based SAFNet [23], and the deep-unfolding-based AFUNet [26]. For fair comparison, we retrain all these methods on our collected dataset using their official open-source implementations. We further compare with existing LDM-based HDR reconstruction methods [12,13,40], which leverage pre-trained LDMs and are fine-tuned on significantly larger datasets than ours. Notably, since UltraFusion [12] is a multi-exposure fusion method that cannot generate HDR images, we only include it for qualitative comparison.

Qualitative Results. We present several challenging scenarios. As shown in Fig. 4, severe camera motion invalidates accurate pixel-wise alignment, leaving only semantic cues for reconstruction. The left case in Fig. 5 involves both human motion and over-saturation, causing severe information loss in the input LDR sequence. In the right example, despite the informative reference frame, camera motion still induces misalignment and degrades reconstruction quality.

As shown in Fig. 4 and 5, existing end-to-end methods consistently suffer from obvious artifacts or ghosting in these challenging regions. Although DiffHDR exploits the generative ability of diffusion models, it still fails to produce visually plausible results due to limited training data. In contrast, our GMODiff combines input cues with powerful priors from pre-trained LDMs to robustly inpaint artifact-prone regions without unrealistic hallucinations, yielding more accurate and perceptually faithful results. We further compare GMODiff with existing LDM-based methods in Fig. 6. GMDD [13] and LE-Diff [40] exhibit severe detail loss because LDMs operate in a compressed latent space. Specifically, GMDD cannot recover missing content in over-exposed areas, while LE-Diff fails to model physical illumination, leading to large brightness and color deviations from the GT. UltraFusion [12] is trained solely on LDR images and thus cannot generate real HDR outputs; it also relies heavily on optical-flow alignment, producing obvious artifacts when alignment fails. By contrast, guided by the regression module, GMODiff strictly follows physical constraints and avoids latent-compression information loss via the degradation-aware decoder, generating more realistic and visually pleasing results.

Qualitative comparisons with Prabhakar *et al.* [33], Sen *et al.* [36], and Tursun *et al.* [39] are provided in the **supplementary material**.

Quantitative Results. Quantitative comparisons on publicly available datasets are summarized in Tab. 1 and Tab. 2. Our method consistently outperforms competing approaches across a diverse set of perceptual metrics. Specifically, it achieves notable gains in both full-reference metrics and no-reference metrics, demonstrating its superior ability to reconstruct visually pleasing and perceptually faithful HDR images. Compared to AFUNet [26], the state-of-the-art deep unfold method, our approach attains competitive performance in distortion-based metrics (PU-PSNR and PU-SSIM) while achieving significantly better perceptual quality-highlighting that leveraging regression priors to guide pre-trained latent diffusion models enables an effective trade-off between fidelity and perceptual realism. Moreover, in contrast to diffusion-based alternatives such as DiffHDR [47], our method fully exploits the rich image priors embedded in large-scale pre-trained LDMs. DiffHDR, trained only on limited in-domain data, lacks robust generic priors and struggles in complex real-world scenarios. Notably, since LE-Diff [40] recalculates physical brightness during the reconstruction process, its HDR results suffer from severe brightness discrepancies, leading to extremely low full-reference metrics. Therefore, we do not report it in the table. The user study is shown in Fig. 8, with details provided in the supplementary material.

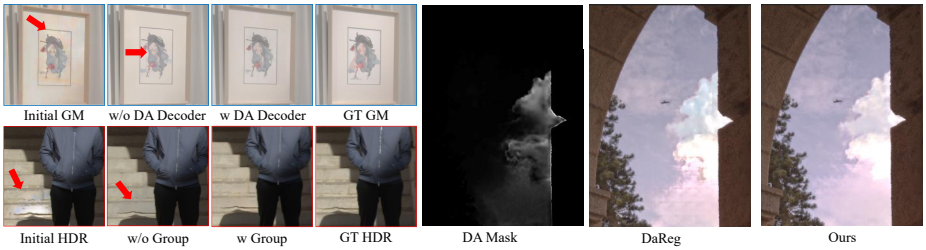

Fig. 7: Qualitative results of our ablation study.

Table 3: Inference time comparison on a 1080 × 1920 image.

Time (s) ↓	AHDR [46]	DiffHDR [47]	AFUNet [23]	SCTNet [38]	UltraFusion [8]	GMDD [13]	LE-Diff [40]	Ours
Inference Time	0.24s	13.52s	6.16s	5.66s	93.4s	39.2s	132.7s	0.93s

4.3 Ablation Studies

To validate the effectiveness of each core component in the proposed GMODiff framework, we conduct comprehensive ablation experiments on the benchmark test set, with quantitative results reported in Tab. 4 and qualitative results presented in Fig. 7. We take the full GMODiff model as the baseline to ablate key modules one by one, and also evaluate the standalone performance of DaReg for comparison. DaReg-Seg refers to the joint training of the reconstruction branch and segmentation head, which leads to a noticeable performance drop. In the second stage of the framework, removing the DA Decoder causes a significant performance decline, as the latent space of LDM fails to preserve the rich detail information contained in the initial GM. Thanks to the rational design of the overall framework, the model only experiences slight performance degradation when other modules are ablated. Compared with the original regression-based DaReg model, GMODiff achieves a favorable trade-off between perceptual quality and distortion metrics with acceptable computational overhead.

4.4 Inference Time Analysis

Tab. 3 compares inference latency on 1920 × 1080 HDR reconstruction. Our method is overall second-fastest, slightly slower than the lightweight CNN-based AHDRNet [46], and significantly faster than transformer-heavy models (AFUNet [26], SCTNet [38]), which, despite using efficient window-based attention, incur high latency due to deep stacking of Transformer blocks and processing at high spatial resolutions. Among diffusion-based approaches, UltraFusion [8] requires over a minute per image, owing to patch-wise processing with 50 denoising steps, optical flow alignment, and ControlNet overhead. DiffHDR [47], despite using only 5 steps, operates in pixel space without latent compression, resulting in high memory bandwidth and slow inference. Since LE-Diff [40] contains two LDMs that separately predict results for different exposures, it takes more than

Table 4: Ablation study of the each component. The best and second best results are colored with red and blue.

Type	PSNR↑	M-IQA↑	MUSIQ↑
DaReg	36.87	0.5819	60.57
DaReg-Seg	35.47	0.5803	59.77
w/o $\mathcal{L}_r$	36.60	0.5875	60.84
w/o Group Training	36.76	0.5858	60.75
w/o DA Mask	36.49	0.5862	60.73
w/o DA Decoder	30.57	0.5754	59.93
GMODiff (ours)	36.68	0.5881	60.96

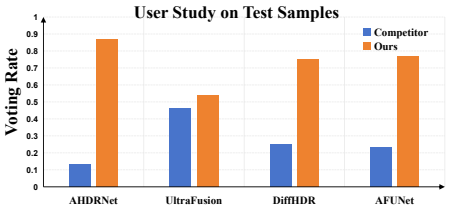

Fig. 8: The user study results on the benchmark test set. Details can be found in the supplementary materials.

2 minutes to infer one complete HDR image. In contrast, our method leverages a pre-trained LDM that performs computation in a highly compressed latent space with a single denoising step, achieving the fastest speed among diffusion-based methods. Moreover, it is compatible with efficient inference frameworks such as `xFormers`, which further reduce latency through memory, efficient attention and kernel fusion, making it practical for real-world applications.

5 Conclusion

We presented GMODiff, a gain map-driven one-step latent diffusion framework for efficient multi-exposure HDR reconstruction. Instead of directly generating HDR radiance in the latent space, GMODiff reformulates the task as degradation-aware gain map refinement within a dual-layer HDR representation, enabling pre-trained LDM priors to be exploited without redesigning the VAE or requiring HDR-specific latent compression. By initializing diffusion from a regression-based gain map estimate and using degradation-aware cues to guide both latent denoising and decoding, the proposed framework combines the structural fidelity of regression models with the perceptual restoration ability of diffusion models while reducing hallucinations in challenging motion and saturated regions. Experiments on benchmark and cross-dataset settings show that GMODiff achieves a favorable balance between distortion fidelity and perceptual quality, with clear improvements in visual realism and artifact suppression.

Acknowledgment

This work is supported by NSFC of China under Grant 62301432 and Grant 62306240, the Natural Science Basic Research Program of Shaanxi under Grant 2023-JC-QN-0685 and Grant QCYRCXM-2023-057, the Fundamental Research Funds for Central Universities, and Guangdong Basic and Applied Basic Research Foundation 2025A1515011119, the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University ZX2025019.

References

1. Adobe: Gain map specification. https://helpx.adobe.com/camera-raw/using/gain-map.html (2024), accessed: 2025-11-13
2. Apple: Explore hdr rendering with edr. https://developer.apple.com/videos/play/wwdc2021/10161/ (2021), accessed: 2025-11-13
3. Azimi, M., et al.: Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In: 2021 Picture Coding Symposium (PCS). pp. 1–5. IEEE (2021)
4. Bogoni, L.: Extending dynamic range of monochrome and color images through fusion. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. vol. 3, pp. 7–12. IEEE (2000)
5. Canham, T.D., Tedla, S., Murdoch, M.J., Brown, M.S.: Gain-mlp: Improving hdr gain map encoding via a lightweight mlp. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18619–18628 (2025)
6. Canham, T.D., Tedla, S.K., Murdoch, M.J., Brown, M.S.: Gamma maps: Nonlinear gain maps for hdr reconstruction. In: London Imaging Meeting. vol. 5, pp. 52–58. Society for Imaging Science and Technology (2024)
7. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European conference on computer vision. pp. 17–33. Springer (2022)
8. Chen, Z., Wang, Y., Cai, X., You, Z., Lu, Z., Zhang, F., Guo, S., Xue, T.: Ultrafusion: Ultra high dynamic imaging using exposure fusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16111–16121 (2025)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS (2021)
10. Du, Y., Zhao, Z., Su, S., Golluri, S., Zheng, H., Yao, R., Wang, C.: Superpc: a single diffusion model for point cloud completion, upsampling, denoising, and colorization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16953–16964 (2025)
11. Gallo, O., Gelfandz, N., Chen, W.C., Tico, M., Pulli, K.: Artifact-free high dynamic range imaging. In: 2009 IEEE International conference on computational photography (ICCP). pp. 1–7. IEEE (2009)
12. Google: Ultra hdr image format. https://developer.android.com/media/platform/hdr-image-format (2024), accessed: 2025-11-13
13. Guan, Y., Xu, R., Liao, Y., Yao, M., Wang, L., Xiong, Z.: Hdr image generation via gain map decomposed diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17536–17545 (2025)
14. Guo, J., Chen, Z., Li, W., Guo, Y., Zhang, Y.: Compression-aware one-step diffusion model for jpeg artifact removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14930–14939 (2025)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
17. Hu, J., Gallo, O., Pulli, K., Sun, X.: HDR deghosting: How to deal with saturation? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1163–1170 (2013)
18. Hu, T., Wu, L., Dong, W., Wu, P., Sun, J., Xu, X., Yan, Q., Zhang, Y.: Boosting hdr image reconstruction via semantic knowledge transfer. IEEE Transactions on Image Processing **35**, 1910–1922 (2026). https://doi.org/10.1109/TIP.2026.3652360

19. Hu, T., Yan, Q., Jin, J., Dong, W., Wu, P., Qi, Y., Lin, W., Zhang, Y.: Ghost-free hdr imaging via latent low-frequency priors and deformable attention alignment. IEEE Transactions on Image Processing **35**, 2684–2698 (2026). https://doi.org/10.1109/TIP.2026.3671656
20. Hu, T., Yan, Q., Qi, Y., Zhang, Y.: Generating content for hdr deghosting from frequency view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25732–25741 (2024)
21. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**(4), 144–1 (2017)
22. Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High dynamic range video. ACM Transactions on Graphics (TOG) **22**(3), 319–325 (2003)
23. Kong, L., Li, B., Xiong, Y., Zhang, H., Gu, H., Chen, J.: Safnet: Selective alignment fusion network for efficient hdr imaging. In: European Conference on Computer Vision. pp. 256–273. Springer (2024)
24. Lan, Y., Cui, Z., Liu, C., Peng, J., Wang, N., Luo, X., Liu, D.: Exploiting diffusion prior for real-world image dehazing with unpaired training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4455–4463 (2025)
25. Lan, Y., Cui, Z., Luo, X., Liu, C., Wang, N., Zhang, M., Su, Y., Liu, D.: When schrodinger bridge meets real-world image dehazing with unpaired training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8756–8765 (2025)
26. Li, X., Ni, Z., Yang, W.: Afunet: Cross-iterative alignment-fusion synergy for hdr reconstruction via deep unfolding paradigm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10666–10675 (2025)
27. Liao, Y., Guan, Y., Xu, R., Li, J., Sun, S., Xiong, Z.: Learning gain map for inverse tone mapping. In: The Thirteenth International Conference on Learning Representations (2025)
28. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: European Conference on computer vision. pp. 344–360. Springer (2022)
29. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in neural information processing systems **35**, 5775–5787 (2022)
30. Ni, Z., Zhang, Y., Ren, K., Yang, W., Wang, H., Kwong, S.: Semantic masking with curriculum learning for robust hdr image reconstruction: Z. ni et al. International Journal of Computer Vision pp. 1–16 (2025)
31. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.: Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. IEEE Transactions on Image Processing **30**, 3885–3896 (2021)
32. Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024)
33. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. In: 2019 IEEE International Conference on Computational Photography (ICCP). pp. 1–8. IEEE (2019)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)

36. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based hdr reconstruction of dynamic scenes. ACM Trans. Graph. **31**(6), 203–1 (2012)
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
38. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12836–12845 (2023)
39. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: An objective deghosting quality metric for HDR images. Comput. Graph. Forum **35**(2), 139–152 (2016)
40. Wang, C., Xia, Z., Leimkuhler, T., Myszkowski, K., Zhang, X.: Lediff: Latent exposure diffusion for hdr generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 453–464 (2025)
41. Wang, L., Yoon, K.J.: Deep learning for hdr imaging: State-of-the-art and future trends. IEEE transactions on pattern analysis and machine intelligence **44**(12), 8874–8895 (2021)
42. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. Advances in Neural Information Processing Systems **37**, 92529–92553 (2024)
43. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 117–132 (2018)
44. Xie, X., Zhang, Q., Zheng, W.S.: Diffusion-based event generation for high-quality image deblurring. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2194–2203 (2025)
45. Yan, Q., Chen, W., Zhang, S., Zhu, Y., Sun, J., Zhang, Y.: A unified hdr imaging method with pixel and patch level. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22211–22220 (2023)
46. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1751–1760 (2019)
47. Yan, Q., Hu, T., Sun, Y., Tang, H., Zhu, Y., Dong, W., Van Gool, L., Zhang, Y.: Toward high-quality hdr deghosting with conditional diffusion models. IEEE Transactions on Circuits and Systems for Video Technology **34**(5), 4011–4026 (2024). https://doi.org/10.1109/TCSVT.2023.3326293
48. Yan, Q., Hu, T., Wu, P., Dai, D., Gu, S., Dong, W., Zhang, Y.: Efficient image enhancement with a diffusion-based frequency prior. IEEE Transactions on Circuits and Systems for Video Technology **35**(9), 8452–8465 (2025). https://doi.org/10.1109/TCSVT.2025.3549351
49. Yan, Q., Yang, K., Hu, T., Chen, G., Dai, K., Wu, P., Ren, W., Zhang, Y.: From dynamic to static: Stepwisely generate hdr image for ghost removal. IEEE Transactions on Circuits and Systems for Video Technology (2024)
50. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. IEEE Transactions on Image Processing **29**, 4308–4322 (2020)
51. Zhou, W., Yang, Y., Hu, T., Hui, P., Jin, J., Cao, Y., Yan, Q., Zhang, Y.: High dynamic range imaging via spatial-frequency interaction. IEEE Transactions on Circuits and Systems for Video Technology (2026)