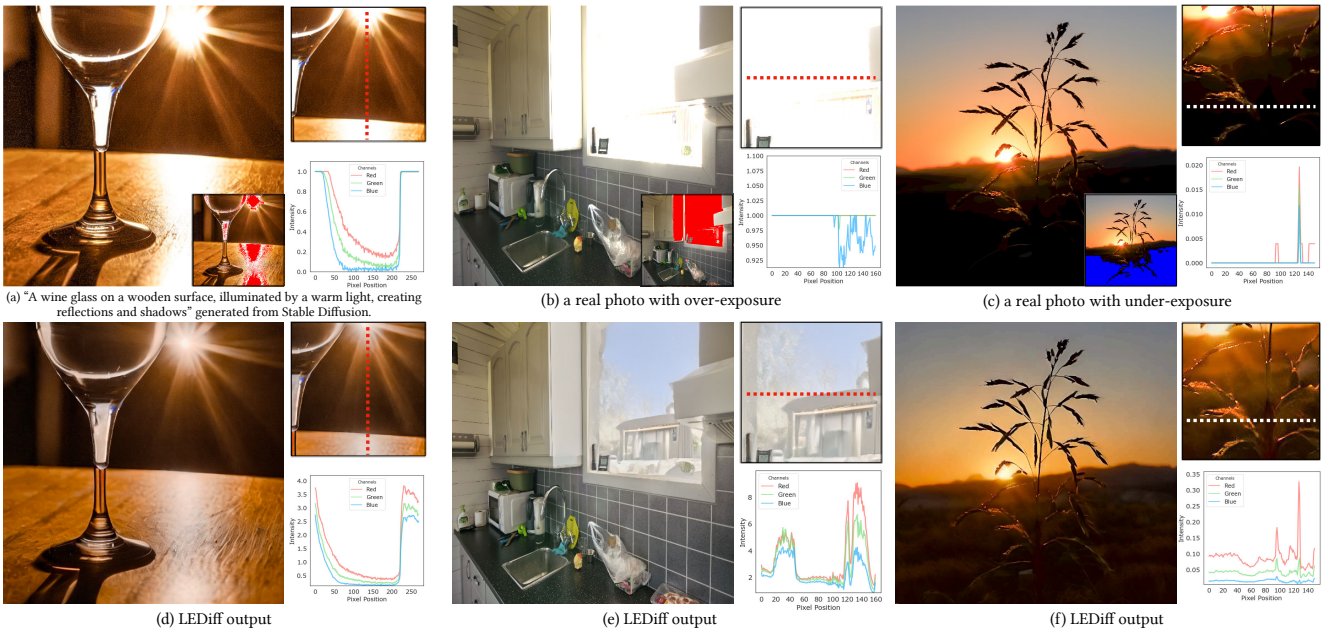
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

LEDiff: Latent Exposure Diffusion for HDR Generation

Chao Wang[1] Zhihao Xia[2] Thomas Leimkühler[1] Karol Myszkowski[1] Xuaner Zhang[2]
[1]MPI Informatik [2]Adobe
{chaowang, karol, thomas.leimkuehler}@mpi-inf.mpg.de {cezhangxer, zhihao.zach.xia}@gmail.com



(a) "A wine glass on a wooden surface, illuminated by a warm light, creating reflections and shadows" generated from Stable Diffusion.

(b) a real photo with over-exposure

(c) a real photo with under-exposure

(d) LEDiff output

(e) LEDiff output

(f) LEDiff output

Figure 1. LEDiff enables high dynamic range (HDR) content generation with photorealistic details in both over- and under-exposed regions by performing exposure fusion in latent space, making it applicable to generated content and real photos mapped to the latent space. While existing generative models (e.g., Stable Diffusion) are restricted to low dynamic range (a) and standard cameras struggle to capture full scene dynamic range, causing clipping in highlights (b) and shadows (c), LEDiff restores both detail and dynamic range (d)–(f), as shown in scanline plots. All HDR images are tone-mapped for visualization and are best viewed on an HDR display. See the supplemental for more details.

Abstract

While consumer displays increasingly support more than 10 stops of dynamic range, most image assets — such as internet photographs and generative AI content — remain limited to 8-bit low dynamic range (LDR), constraining their utility across high dynamic range (HDR) applications. Currently, no generative model can produce high-bit, high-dynamic range content in a generalizable way. Existing LDR-to-HDR conversion methods often struggle to produce photorealistic details and physically-plausible dynamic range in the clipped areas. We introduce LEDiff, a method that enables a generative model with HDR content generation through latent space fusion inspired by image-space exposure fusion techniques. It also functions as an LDR-to-HDR converter, expanding the dynamic range of existing low-dynamic range images. Our approach uses a small HDR dataset to enable a pretrained diffusion model to recover detail and dynamic range in clipped highlights and shadows. LEDiff brings HDR capabilities to existing generative models and converts any LDR image to HDR, creating photorealistic HDR outputs for image generation, image-based lighting (HDR environment map generation), and photographic effects such as depth of field simulation, where linear HDR data is essential for realistic quality.

1. Introduction

Generative models today can produce highly realistic and creative visual content, yet they remain largely confined

to 8-bit, low dynamic range (LDR) representations. This limitation causes bright highlights and deep shadows to be clipped, preventing the full capture of light and color variation present in real-world scenes. As a result, these models restrict further photographic editing, as stronger tone mapping cannot reveal any additional details. Additionally, LDR limits downstream applications, such as image-based lighting and depth of field rendering, where realistic visual quality depends on HDR data. The need for HDR content has also grown with the rise of consumer displays that support HDR, often with over 10 stops of dynamic range, making the screen feel like a window looking out at the real scene [8, 12]. The longstanding popularity of LDR content — optimized for traditional display purposes through clipping, gamma correction, and quantization — may no longer hold as HDR displays become increasingly prevalent.

HDR reconstruction has been widely studied in recent years and generally falls into two main categories. The first approach focuses on multi-exposure fusion, which assumes a set of exposure-bracketed photos as input and primarily addresses alignment issues among these images [13, 16, 28, 33, 38, 47, 49, 56, 57, 60, 63, 65, 67–70]. The second approach centers on inverting the tone mapping applied to the input LDR image, aiming to recover missing details in clipped regions [20, 21, 34, 37, 41, 55, 71, 75]. With the growth of generative models, approaches to HDR generation also emerged. GlowGAN [62] is the first attempt to generate HDR content from scratch and can extend existing LDR images through GAN inversion. However, its effectiveness is limited by the class-specific training and generation constraints inherent to GANs.

Our goal is to repurpose a pre-trained latent diffusion model (e.g., Stable Diffusion [53]) to generate a closely related yet distinct distribution: linear HDR content. To maintain the generative capabilities of the pre-trained model without compromising it by training on limited HDR data, we aim to retain the powerful latent space of the original latent diffusion model. Enabling HDR capabilities in a pre-trained generative model requires 1) hallucinating clipped area in highlights and shadows, 2) linearizing the image and expanding its dynamic range.

Interestingly, we observe that the latent space of a diffusion model strongly correlates with the image space in terms of clipping and pixel intensity; pixels that are clipped in the image space are similarly clipped in the latent space. This motivates our approach of aiming for a clipping-free latent space. We then fine-tune the decoder to linearize and expand the dynamic range, generating an HDR image in the image space. To get a clipping-free latent code, we draw inspiration from image-space exposure fusion, applying this concept in latent space by merging latent code that represents different exposure-level captures. To generate these bracketed codes, we train a highlight generator that takes as input a latent code with highlight clipping (representing an over-exposed image) and produces a lower-exposed latent code without clipping. Similarly, we train a shadow generator that takes an under-exposed latent code as input and generates a higher-exposed latent code free from shadow clipping. We train the highlight and shadow generators using a relatively small HDR dataset, relying on the preserved latent space to offload intensive generation tasks to the pre-trained model. Once we have these exposure-bracketed latent codes, a learnable fusion module merges them into a single, clipping-free latent code.

In summary, our approach involves training highlight and shadow generators to hallucinate missing details in over- and under-exposed areas of the latent space and fine-tuning the decoder to linearize and expand dynamic range during decoding.

This approach offers several advantages. Leveraging the *original pre-trained latent space* allows our method to act as a plug-and-play solution, converting any latent code to HDR, whether for generated content or real images encoded into the latent space. Using a *learnable latent space exposure fusion*, combined with a *learnable linearization decoder*, eliminates the need for hand-crafted merging features [44] and avoids the necessity of estimating exposure parameters and camera response curves [7] required by traditional exposure fusion algorithms.

Our method equips a generative model with HDR generation capabilities, greatly expanding its applicability across various domains. We demonstrate our method using the stable diffusion model [53]. The main contributions of this work are as follows:

- A fine-tuning approach that fully preserves a pre-trained (LDR) latent space and uses a small set of HDR data to expand dynamic range and add details to both deep shadows and bright highlights.
- A method that enables HDR generation in existing image and video generative models and achieves state-of-the-art performance in inverse tone mapping for real LDR images, including linearization and photorealistic reconstruction in clipped regions.

2. Related Work

2.1. Generative Image Models

The last decades have witnessed the rapid development of generative image models. In the early stage, GAN-based [29, 30, 35, 59, 66, 73, 74, 77–79] methods dominate this domain and focus on domain-specific generations.

Autoregressive Transformer models, including DALL·E [50], CogView [18], and PARTI [72], achieve high-quality text-to-image generation by encoding images as discrete tokens and learning text-image distributions from large datasets. However, their sequential generation process re-

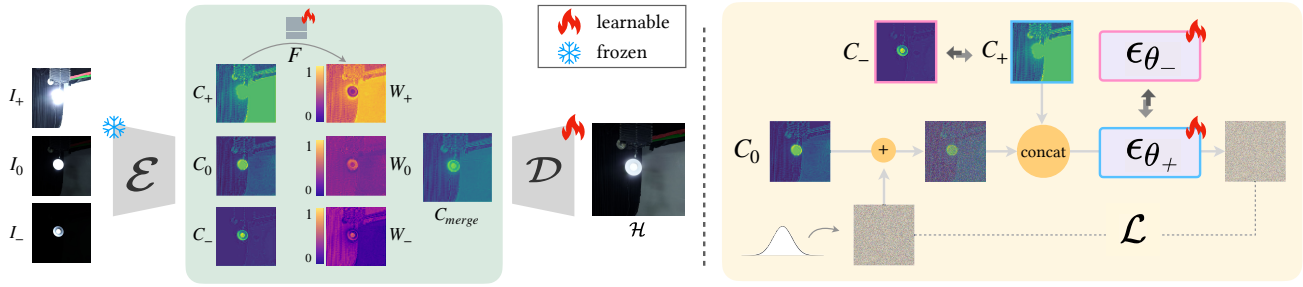


Figure 2. Finetuning scheme of decoder and denoiser. (Left: fine-tuning the decoder) Exposure bracketed images I_+, I_0, I_- are encoded via the pre-trained encoder to generate corresponding latent codes. These latent codes are fused using a learnable fusion module $\mathcal{F}$ to produce a latent code $\mathcal{C}_{\text{merge}}$ free of clipping, which is then decoded into an HDR image $\mathcal{H}$ through the finetuned decoder. (Right: fine-tuning the denoiser) The model processes the latent code, either sampled from noise or encoded from the image. It utilizes $\mathcal{C}_+$ to train the highlight denoiser ϵ_{θ_-} and $\mathcal{C}_-$ to train the shadow denoiser ϵ_{θ_+}, while also incorporating $\mathcal{C}_0$ corrupted by randomly sampled noise.

sults in slower image synthesis.

Diffusion-based text-to-image (T2I) models have advanced rapidly, each iteration pushing the boundaries of efficiency and quality. GLIDE [46] introduced Gaussian noise combined with CLIP-encoded text, enabling both classifier-guided and classifier-free T2I generation. Imagen [54] enhanced fidelity by using large language models for text encoding and scaling images through super-resolution layers. DALLE-2 [51] linked text and image latents with a CLIP-based prior model, predicting denoised images across time steps. Stable Diffusion [53] improved efficiency by mapping images to a compact latent space, accelerating the diffusion process, while VQ-Diffusion [25] leveraged VQ-VAE [61] to conduct discrete diffusion in a compressed space, producing high-quality outputs with reduced computational demands.

2.2. High Dynamic Range Imaging

High Dynamic Range (HDR) imaging has advanced significantly over recent decades. HDR images are often reconstructed from multiple-exposure LDR images. Debevec et al. [16] achieve this by capturing static images at different exposures, estimating the camera response, and merging them. To handle motion in dynamic scenes, Sen et al. [56] introduce a patch-based method, while Kalantari et al. [28] bring this into the deep learning era with a dataset for training an end-to-end HDR network. Later works build on this by using advanced architectures like U-Net [63] and transformers [13, 38, 57, 60], and by integrating attention, selective modules, and generative priors for enhanced HDR performance [33, 67–70].

An alternative approach to acquiring HDR is inverse tone mapping (ITM), where HDR content is generated from a single LDR image, addressing linearization, dynamic range extension, and hallucination [6]. Early ITM methods focused on linearization and dynamic range extension for display quality improvement [2, 4, 5, 17, 42, 43]. Eilertsen et al. [20] introduced deep learning to ITM using CNNs to enhance highlight detail, inspiring later works that improve artifact reduction, linearization, and overexposure hallucination with multi-branch networks, mask mechanisms, and attention modules [37, 41, 55, 71].

Another ITM direction generates multi-exposure brackets from a single image, merging them via traditional methods. Techniques include exposure generation with 3D U-Nets [21], recursive networks [34], and adaptive networks for enhanced fusion [75]. However, these methods often produce blurred results due to their mean-seeking nature.

More recently, Wang et al. [62] modeled HDR-LDR relationships using Gaussian exposure assumptions, enabling unlabeled HDR generation and improved over-exposed region reconstruction via pre-trained models. Concurrent works [7, 24] leverage diffusion models for reconstructing HDR, creating exposure brackets without tuning, though they rely on merging to finalize HDR output. Our approach differs by performing exposure merging in the latent space, eliminating exposure parameter estimation and is able to achieve greater dynamic range.

3. Method

3.1. Preliminaries

A latent diffusion model (LDM), such as Stable Diffusion [53], consists of a variational autoencoder (VAE), a denoising UNet, and a text encoder. The VAE encoder $\mathcal{E}$ compresses an image into a low-dimensional latent space. The denoising UNet then iteratively refines this noisy latent code into its clean version, which is subsequently decoded by the VAE decoder $\mathcal{D}$ into an LDR image.

LDMs are inherently limited to generating LDR images for two reasons. First, the VAE is trained exclusively on LDR images, constraining its ability to represent the dynamic range for HDR content. Second, the denoising UNet is designed to produce latent codes that encode only LDR

images. These limitations are demonstrated in Fig. 3. While diffusion models can plausibly hallucinate details in clipped regions — similar to their performance in inpainting tasks [15, 39, 64] — the challenge lies in guiding them to generate images with true high dynamic range and to preserve details in both highlights and shadows.

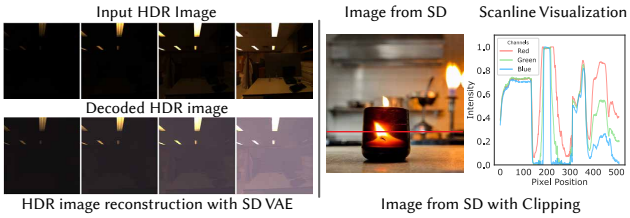


Figure 3. Limitations of the vanilla Stable Diffusion (SD) in generating HDR content. **Left:** The limitation of the SD VAE in encoding and decoding an HDR image, visualized in multiple exposure levels, which reveals a significant fidelity loss, especially in the shadow. **Right:** An image generated with SD, along with a scanline that shows pixel clipping in highlight and shadow regions.

A straightforward approach to adapting LDMs for HDR generation is to fine-tune both the VAE and the UNet with HDR data. However, this requires a significant amount of HDR data and does not fully utilize the generative priors of the pre-trained model. Instead, we propose to preserve the latent space of a pre-trained latent diffusion model as is, generating LDR latent codes at multiple exposure levels — a "latent exposure bracket" consisting of shorter and longer exposure variants. These latent codes are then fused in the latent space using a learnable lightweight fusion module. This merged HDR latent code is subsequently decoded by a fine-tuned VAE decoder to generate an HDR image.

In Sec. 3.2, we describe merging exposure brackets within the latent space and fine-tuning a specialized HDR decoder. In Sec. 3.3, we outline the approach for generating the latent exposure brackets.

3.2. Merging Latent Exposure Brackets

A key idea in our method is to preserve the priors learned by a pre-trained latent diffusion model by restricting it to generating latent codes for LDR images – the domain in which the model has been extensively trained on using billions of samples. The main challenge, then, is to effectively merge these LDR latent codes and decode them into an HDR image. To achieve this, we propose merging the latent LDR exposure bracket directly in the latent space and only fine-tuning the VAE decoder on a small set of HDR images to reconstruct the final HDR output, as shown in Fig. 2-left.

We simulate exposure bracketed images from real HDR images. For each HDR image $\mathcal{H}$, we use the method from [3] to calculate the lower (E_-) and upper (E_+) exposure bounds needed to cover the entire scene's dynamic range. We then project $\mathcal{H}$ into three exposure intervals, $\{E_-, E_0, E_+\}$, where $E_0 = (E_- + E_+)/2$, and randomly sample camera response curves (CRFs) from [20] to introduce non-linearity. This generates an exposure bracket of low-, medium-, and high-exposure LDR images: I_-, I_0, I_+, as follows:

$$I_i = \frac{(1+\beta)\min(E_i, 1.0)^\gamma}{\beta + \min(E_i, 1.0)^\gamma}, \tag{1}$$

with $\beta \sim \mathcal{N}(0.6, 0.1)$ and $\gamma \sim \mathcal{N}(0.9, 0.1)$. We incorporate this randomization into the CRF sampling to improve the training data diversity. We use the pre-trained encoder $\mathcal{E}$ to map these LDR images into the latent codes: C_-, C_0, C_+. Further details can be found in Sec. 4.

Visualizing these latent codes (Fig. 2) reveals a strong correlation between the latent space and image space, where intensity levels in over- and under-exposed regions are reflected similarly in the latent space. Thus, analogous to image-space HDR merging techniques that rely on brightness, color, and saturation to identify valid pixels [44], we hypothesize that an efficient merging of the latent exposure bracket can be achieved by deriving a per-pixel weight map for each channel based on local regions of the latent codes.

To implement this, we introduce a straightforward yet effective learnable fusion module, $\mathcal{F}$. Each latent code is processed through a depth-wise convolution [14] to produce an initial weight map, which is then normalized using a softmax function to ensure the merging weights sum to one across each pixel and channel. This merged latent code C_{merge}, combining information from multiple exposures, retains scene details without clipping. We then decode C_{merge} into an HDR image in log space, fine-tuning the decoder with the original VAE's loss functions [53], including both a reconstruction loss and a GAN loss that compares $\mathcal{D}(C_{\text{merge}})$ with $\mathcal{H}$ in log space:

$$\mathcal{L}_{\mathcal{F},\mathcal{D}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{GAN}} \tag{2}$$

3.3. Generating Latent Exposure Brackets

With a trained $\mathcal{F}$ and a fine-tuned $\mathcal{D}$, the remaining step is to generate an exposure bracket in the latent space, Fig. 2-right.

We assume a single latent code is given, which may originate from either 1) a real image encoded into the latent space or 2) a text prompt. This allows LEDiff to naturally handle generated contents (i.e., Text-to-HDR generation) and real photos (i.e., LDR-to-HDR conversion).

When given a single LDR latent code C, we need to expand it into an exposure bracket $\{C_-, C_0, C_+\}$, which involves hallucinating details in over-exposed highlights and under-exposed shadows. We handle the hallucination of highlights and shadows separately but with similar processes, which is inspired by the conditioning strategy of [31]. For simplicity, we describe the highlight hallucination process: generating a lower-exposure latent code, C_-

from C_0, and C_0 from C_+. This method implicitly learns to reduce the exposure of the conditioned image, effectively restoring details in clipped highlights.

Denote the highlight hallucination denoiser as $\epsilon_{\theta_-}(\cdot)$ with parameters θ_-. We fine-tune $\epsilon_{\theta_-}(\cdot)$ from the pre-trained Stable Diffusion denoiser to generate C_- conditioned on the input latent C_0 and generate C_0 conditioned on the input latent C_+. The conditioning is achieved by concatenating the input condition with the noisy input before feeding it into the denoiser, with a modified first layer to handle change in channels. The rest of the network remains initialized with the pretrained weights. Fine-tuning uses a standard denoising diffusion loss.

To train $\epsilon_{\theta_-}(\cdot)$, we first randomly sample an exposure bracket from the HDR dataset, following the process outlined in Sec 3.2. These samples are encoded using the pre-trained VAE encoder, producing a latent exposure bracket $\{C_-, C_0, C_+\}$. We then corrupt the underexposed latent C_- and C_0 with Gaussian noise at a randomly sampled timestamp t. The task of the denoising network, $\epsilon_{\theta_-}(\cdot)$ is to predict the noise $\hat{\epsilon}_{C_-,C_0} = \epsilon_{\theta_-}(C_0, C_-^t, t)$ and $\hat{\epsilon}_{C_0,C_+} = \epsilon_{\theta_-}(C_+, C_0^t, t)$ using the objective:

$$\mathcal{L}_C = \mathbb{E}_{C,\epsilon_{C_-,C_0}\sim\mathcal{N}(0,\mathcal{I}),t\sim\mathcal{U}(T)} \left\|\epsilon_{C_-,C_0} - \hat{\epsilon}_{C_-,C_0}\right\|_2^2 \quad (3)$$

$$+ \mathbb{E}_{C,\epsilon_{C_0,C_+}\sim\mathcal{N}(0,\mathcal{I}),t\sim\mathcal{U}(T)} \left\|\epsilon_{C_0,C_+} - \hat{\epsilon}_{C_0,C_+}\right\|_2^2 \quad (4)$$

During inference, given an LDR image, we iteratively apply the fine-tuned denoiser to generate its corresponding lower-exposure latent codes.

For generating a higher-exposure latent code C_+, C_0 that hallucinates shadow details, we follow the same procedure to fine-tune a conditional denoiser $\epsilon_{\theta_+}(\cdot)$ with parameters θ_+, conditioned on C_0 and C_- respectively.

After generating the latent exposure bracket, we merge it using $\mathcal{F}$, then decode the merged latent code into an HDR image with the fine-tuned HDR decoder.

4. Dataset and Training

Datasets We gather HDR images from multiple sources [1, 10, 22, 23, 28, 37, 48, 60]. For the panorama datasets [1, 10, 23], we create regular images by projecting the panoramas from random camera viewpoints. In total, we obtain 36,000 HDR images, and then apply the method mentioned in Sec. 3.2 to generate the training pairs $\{I_-, I_0, I_+\}$ and $\{\mathcal{H}\}$.

Training We use this dataset to train $\mathcal{F}$ and to fine-tune $\mathcal{D}$, $\epsilon_{\theta_-}(\cdot)$ and $\epsilon_{\theta_+}(\cdot)$. $\mathcal{F}$ and $\mathcal{D}$ are trained with $\{I_-, I_0, I_+\}$ as inputs and $\{\mathcal{H}\}$ as the ground truth. For $\epsilon_{\theta_-}(\cdot)$ training, we use the pairs $\{I_0, I_+\}$ and $\{I_-, I_0\}$, where the first image serves as the conditioning input and the second is used as the input (to be applied with noise) and ground truth. Similarly, for $\epsilon_{\theta_+}(\cdot)$ training, we use the pairs $\{I_0, I_-\}$ and $\{I_+, I_0\}$. We fine-tune the VAE decoder for 200,000 steps with a learning rate of 10^{-6} and the denoiser for 400,000 steps with a learning rate of 10^{-5}. All fine-tuning processes use the Adam optimizer [32].

5. Applications and Evaluations

We evaluate LEDiff across various HDR content generation tasks: text-to-image, text-to-panorama, and image-to-video (Sec. 5.1), and use it for HDR reconstruction from an LDR image (inverse tone mapping) in Sec. 5.2. Our results cover diverse scenes, demonstrating the generalization ability of LEDiff and its seamless integration into applications benefiting from HDR capabilities. For more results, please refer to the supplementary material.

5.1. HDR Content Generation

Text to HDR Image We compare the text-to-image of LEDiff with that of Stable Diffusion (SD). Starting with a generated latent code C_+, we apply diffusion with the denoiser ϵ_{θ_-} to generate two lower exposure latent codes, C_0 and C_-. The fusion module $\mathcal{F}$ merges these latent codes into $C_{\text{merge}} = \mathcal{F}(C_-, C_0, C_+)$, which is then decoded into an HDR image via $\mathcal{D}$. Fig. 4-left shows that LEDiff achieves a higher dynamic range without clipping, while SD output suffers from clipping. We also demonstrate the effectiveness of HDR for further image editing, such as applying synthetic defocus, where linear HDR content is crucial for realistic depth of field rendering, especially for bokeh simulation [45, 76].

Text to HDR Panorama Our method seamlessly integrates with existing SD-based panorama generation models. Using MVDiffusion [58] as the baseline, LEDiff enables HDR panoramas, which are essential for image-based lighting. As shown in Fig. 4-right, using the HDR panorama as the environment light map enhances contrast and produces more realistic highlights in the rendered image.

Image to HDR Video Our method can be incorporated into SD-based image-to-video models [9], facilitating the generation of HDR video from a single LDR image. Examples of this application are included in the supplemental.

5.2. LDR to HDR Reconstruction

Another promising application of LEDiff is inverse tone mapping. Given an LDR image I_+ with potential clipping, we obtain its latent code $C_+ = \mathcal{E}(I_+)$. Next, we generate lower exposure latents C_0 and C_-, which are fused into a merged latent code $C_{\text{merge}} = \mathcal{F}(C_-, C_0, C_+)$. Finally, the HDR reconstruction $\hat{\mathcal{H}}$ is obtained through the decoder $\mathcal{D}$.

We evaluate our method using standard image quality metrics. However, most established metrics are optimized for LDR content. To better evaluate the HDR outputs, we also conduct a user study in a controlled environment, displaying the content on an HDR display to assess human perception of the HDR contents.

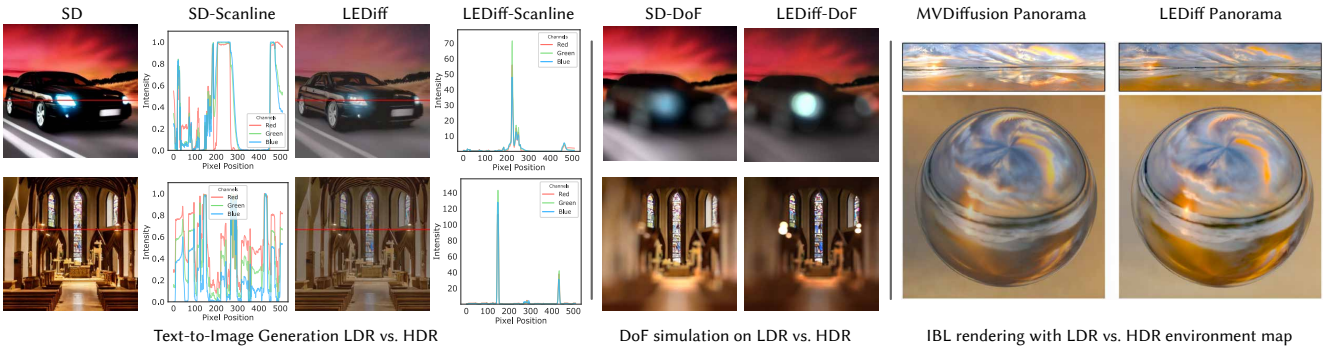


Figure 4. **Left:** Text to HDR image generation using prompts: (1) **"Bright car headlights on a narrow street at night"** and (2) **"A grand church interior with tall stained glass windows, intricate wooden arches, and warm lighting"**. We compare the images generated by SD and LEDiff. We plot a scanline passing through the car headlights and the bright luminaries to demonstrate our ability to produce a wide dynamic range with photorealistic detail. We also show an application of synthetic depth of field (**DoF**), where a linear HDR image is crucial for rendering realistic defocus effects. **Right:** Comparisons are made between the panoramas produced by MVDiffusion [58] and our approach with a prompt **"A peaceful beach at sunset with soft clouds in the sky"**, along with their respective image-based lighting results. Our method leads to higher contrast and an overall more realistic appearance.

We compare our method with five inverse tone mapping methods: **HDRCNN** [20], **MaskHDR** [55], **SingleHDR** [37], **ExpandNet** [41], and **GlowGAN** [62]. Since **GlowGAN** handles only class-specific inputs, we quantitatively evaluate it alongside our method on a subset of **SI-HDR**[26], but exclude it from the user study.

Image Quality Metrics In our quantitative evaluation, we evaluate on images from **SI-HDR** [26], which are excluded from the training dataset. We use the full-reference HDR-VDP3 [40] and no-reference PU21-PIQE [26] metrics for quality evaluation. Since standard metrics struggle to assess hallucinated details in clipped regions, we also use the FID score [27] to evaluate image distributions. For HDR evaluation with FID, we tone-map both generated and ground-truth images to the LDR domain using Reinhard [52], Durand [19], and Liang [36]. Following Chai et al. [11], we generate 60 random 128×128 pixel crops per image, resulting in $10k$ patches for FID computation. While tone-mapped images have different feature statistics from typical LDR images, we found that our FID scores correlate well with perceived quality and maintain consistent rankings across tone-mapping methods.

Previous works [20, 37, 55] primarily focus on hallucinating highlight regions, with limited attention to shadows. By contrast, our approach enables effective hallucination across both highlight and shadow areas. For highlight hallucination, we follow [20, 55, 62] to blend the generated content with the inputs. The process models HDR luminance in the non-overexposed regions of the LDR image, aligning it with the corresponding HDR area. More details are provided in the supplemental. Fig. 5 shows that our method produces more natural hallucinated content within clipped highlight and shadow regions, while other methods produce low-quality artifacts and blur. This is supported by the quantitative results in Table 1.

User Study To comprehensively evaluate the quality of our generated results for the ITM task, we conducted a subjective study. The evaluation included 60 scenes, with half drawn from the **SI-HDR** dataset and the remaining half comprising LDR images collected from online sources or real-world captures. For each scene, a competing method was randomly selected and compared to our approach, with participants tasked to express their preferences. The study was carried out using an HDR display (Dell UP3221Q, 3840×2160 resolution) in a standard office environment with natural lighting, with participants positioned 0.5 meters from the screen. The study engaged 20 participants, all possessing normal or corrected-to-normal vision, resulting in a total of 1200 pairwise comparisons. The outcome of this assessment is detailed in Table 2. The results indicate that our method significantly outperforms other approaches, demonstrating a substantial margin of superiority.

6. Ablation Study

We conduct the following experiments to illustrate the contribution of each component in our method, by excluding: 1) VAE decoder finetuning and 2) denoiser finetuning. We also use SD-based inpainting as an alternative approach for hallucinating clipped areas. We follow Sec. 5.2 for evaluation. Since the inpainting needs an extra mask as input, we extract the clipped region with a threshold.

Results are shown in Fig. 6, with quantitative evaluation in Table 3. As illustrated in Fig. 6, omitting the decoder leads to hallucinations in overexposed regions, but restricts the content to low dynamic range (LDR). In contrast, without denoiser fine-tuning, the dynamic range is extended, but hallucinations within clipped regions are not achieved.

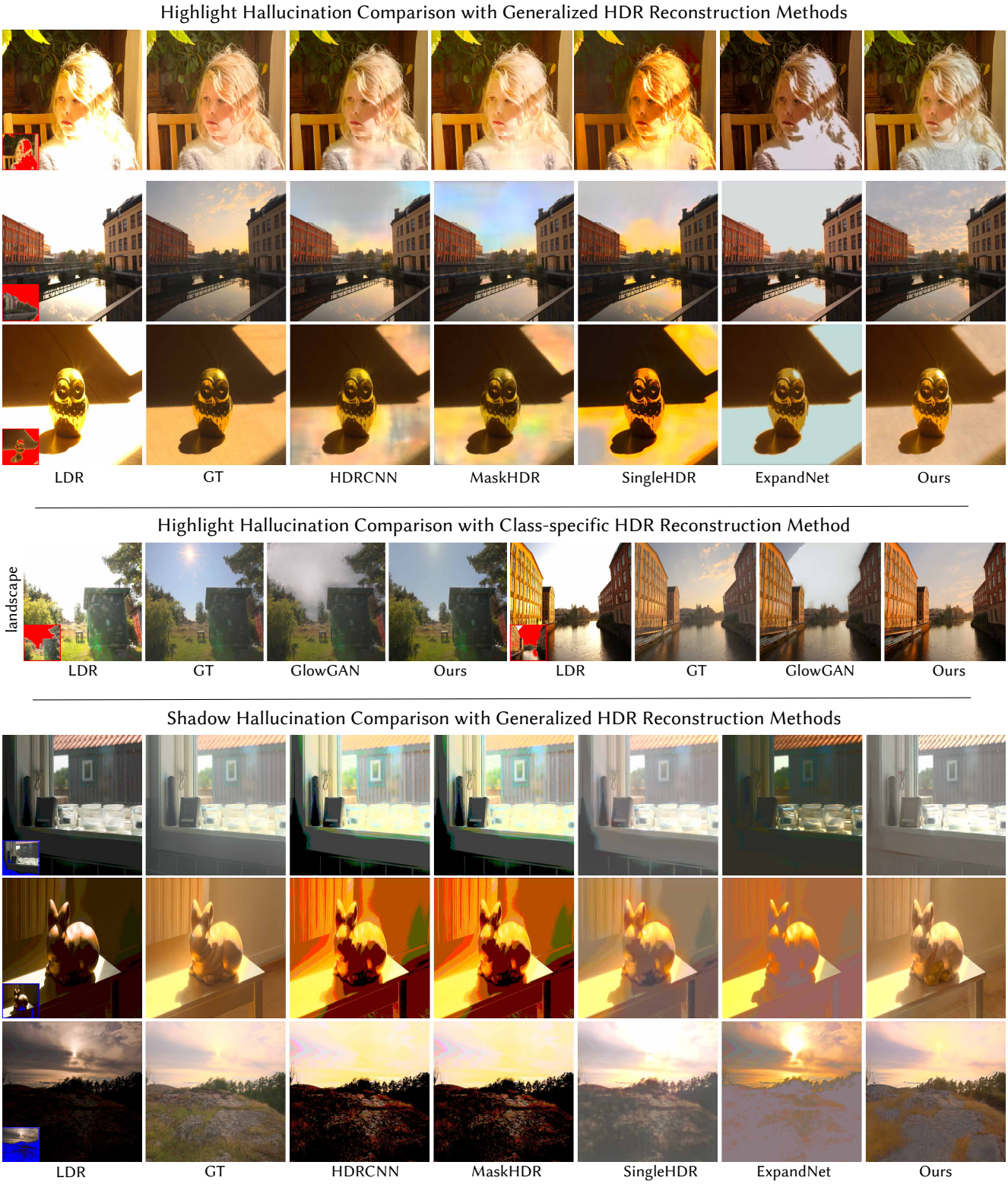


Figure 5. LDR-to-HDR image reconstruction comparisons. Our method effectively hallucinates details in both over- and under-exposed regions, while previous approaches [20, 37, 41, 55, 62] struggle to produce plausible results, especially in shadow regions that they do not address (e.g., HDRCNN and MaskHDR yield identical results for shadow hallucination, as both methods process non-clipped regions in the same way.). Images are tone-mapped for visualization. Best viewed in HDR on an HDR display; see the supplemental for further details.

Method	Highlights					Shadow				
	HDR-VDP3 ↑	PU21-PIQE ↓	FID-R ↓	FID-D↓	FID-L↓	HDR-VDP3 ↑	PU21-PIQE ↓	FID-R ↓	FID-D↓	FID-L↓
HDRCNN [20]	6.90 ± 1.10	49.43 ± 6.91	13.39	16.95	16.67	6.56 ± 1.10	70.95 ± 7.84	32.55	44.20	26.41
MaskHDR [55]	6.47 ± 1.06	49.38 ± 6.95	12.83	13.85	15.21	6.49 ± 1.24	70.98 ± 8.64	32.54	44.43	26.10
SingleHDR [37]	6.13 ± 0.87	49.74 ± 7.29	28.68	34.53	29.50	7.47 ± 0.78	46.04 ± 9.41	39.99	57.28	37.17
ExpandNet [41]	5.23 ± 1.61	52.53 ± 6.50	18.85	25.49	21.34	5.40 ± 0.93	60.01 ± 8.37	32.60	45.50	29.10
Ours	6.16 ± 0.97	48.46 ± 7.04	12.70	13.08	13.73	6.98 ± 0.83	43.37 ± 8.04	20.93	25.54	24.26
GlowGAN [62]	5.67 ± 1.24	46.18 ± 8.22	27.44	40.43	43.85	–	–	–	–	–
Ours∗	5.92 ± 1.17	45.29 ± 7.27	15.62	17.89	16.19	–	–	–	–	–

Table 1. Quantitative evaluation on content hallucination (Ours∗ denotes the test results of our method on a subset of the test set, selected to meet GlowGAN's requirement for class-specific inputs). Full-reference metrics are less suited to generative models due to natural deviations in hallucinated content from the original HDR scene. Our method achieves competitive performance on the full-reference HDR-VDP3 metric and outperforms alternative methods on the no-reference PU21-PIQE metric, which better captures image naturalness and perceptual fidelity. To complement these metrics, we use FID [27] to evaluate the distribution of generated and ground-truth HDR images. Since FID is optimized for LDR, we tone map all HDR images using three tone mapping operators, resulting in three FID scores: FID-R [52], FID-D [19], and FID-L [36]. Across all tone-mapping methods, our method consistently achieves lower FID scores, demonstrating closer alignment with the HDR ground-truth distribution and superior overall quality.

Total	Ours	HDRCNN	Total	Ours	MaskHDR
310	84.19%	15.81%	297	88.22%	11.78%
Total	**Ours**	**SingleHDR**	**Total**	**Ours**	**ExpandNet**
283	89.40%	10.60%	310	94.52%	5.48%

Table 2. User study comparisons with other methods. Item **Total** shows the total number of comparisons, item **Ours** reflects the percentage of users who preferred our results, and the other item indicates the percentage of users who chose other methods. In all cases, we observe a statistically significant preference for our method, as confirmed by a binomial test ($p < 0.01$).

Additionally, using an SD-based inpainting method proves suboptimal, as it relies on irregular masks that differ significantly from those seen during training, resulting in unnatural results.

We choose to use an exposure bracket with three exposure levels, as it is a common practice in image-based exposure fusion [13, 28, 38, 57, 60, 63]. We explore the impact of using two and five exposure levels, detailed in the supplemental.

Method	HDR-VDP3 ↑	PU21-PIQE ↓
wo VAE	4.67 ± 0.87	48.60 ± 7.02
wo denoiser	5.59 ± 1.12	50.25 ± 6.55
inpainting	3.65 ± 1.20	50.18 ± 7.53
Ours	6.16 ± 0.97	48.46 ± 7.04

Table 3. Ablation study on individual components. The table shows that fine-tuning both the VAE decoder and the denoiser yields the highest performance.

7. Conclusion and Limitations

We propose LEDiff, a method for HDR content generation that captures the full dynamic range of real-world scenes. LEDiff operates within the latent space of a pre-trained diffusion model, leveraging its generative capabilities while enabling HDR by reconstructing details in over- and under-exposed regions and performing dynamic range extension. This is accomplished by fine-tuning the VAE decoder and denoiser on a relatively small HDR dataset. LEDiff is agnostic to how the LDR latent code is generated, making it adaptable as an LDR-to-HDR converter for any LDR image encoded into latent space. LEDiff unlocks a range of HDR applications, including HDR panorama generation for realistic image-based lighting and HDR video generation.

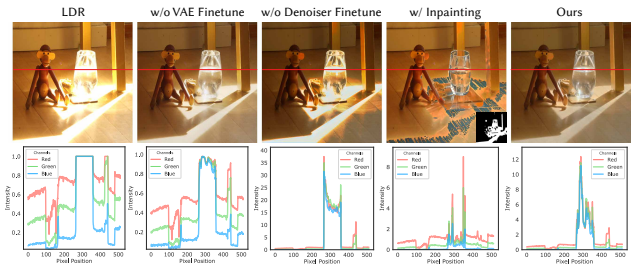


Figure 6. Ablation study. The first row displays the input LDR image alongside the reconstructed outputs; the second row plots the scanline. The mask used for inpainting is shown in the inset. Fine-tuning both the decoder and denoiser is essential for achieving hallucination and dynamic range extension.

However, our approach has limitations. Fine-tuning Stable Diffusion constrains LEDiff to SD's strengths and weaknesses, while using Stable-Video-Diffusion (SVD) for HDR video generation introduces more artifacts due to its weaker generative power. Additionally, our method does not yet handle degradations like compression artifacts (e.g., ringing and blocking) or noise in the input LDR, which could improve robustness to real-world images. Addressing these limitations could improve generalization and is a promising direction for future work.

As HDR displays become more widely available on consumer devices, we anticipate that the demand for HDR content will continue to grow, and believe that LEDiff's potential to expand HDR content creation will help meet the increasing demand for photorealistic HDR contents.

References

[1] Poly haven. 2024. Accessed: 2024-10-21. 5

[2] Ahmet Oğuz Akyüz, Roland Fleming, Bernhard E Riecke, Erik Reinhard, and Heinrich H Bülthoff. Do hdr displays support ldr content? a psychophysical evaluation. *ACM Transactions on Graphics (TOG)*, 26(3):38–es, 2007. 3

[3] Pontus Andersson, Jim Nilsson, Peter Shirley, and Tomas Akenine-Möller. Visualizing errors in rendered high dynamic range images. In *Eurographics 2021 - Short Papers*. The Eurographics Association, 2021. 4

[4] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Expanding low dynamic range videos for high dynamic range applications. In *Proceedings of the 24th Spring Conference on Computer Graphics*, pages 33–41, 2008. 3

[5] Francesco Banterle, Patrick Ledda, Kurt Debattista, Marina Bloj, Alessandro Artusi, and Alan Chalmers. A psychophysical evaluation of inverse tone mapping techniques. In *Computer Graphics Forum*, pages 13–25. Wiley Online Library, 2009. 3

[6] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced high dynamic range imaging*. AK Peters/CRC Press, 2017. 3

[7] Mojtaba Bemana, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Exposure diffusion: Hdr image generation by consistent ldr denoising. *arXiv preprint arXiv:2405.14304*, 2024. 2, 3

[8] Greg Benz. Create and edit true HDR (High Dynamic Range) images - Greg Benz Photography, 2022. 2

[9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 5

[10] Christophe Bolduc, Justine Giroux, Marc Hébert, Claude Demers, and Jean-François Lalonde. Beyond the pixel: a photometrically calibrated hdr dataset for luminance and color prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8071–8081, 2023. 5

[11] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *European Conference on Computer Vision*, pages 170–188. Springer, 2022. 6

[12] Eric Chan. High dynamic range explained, 2023. 2

[13] Rufeng Chen, Bolun Zheng, Hua Zhang, Quan Chen, Chenggang Yan, Gregory Slabaugh, and Shanxin Yuan. Improving dynamic hdr imaging with fusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 340–349, 2023. 2, 3, 8

[14] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 4

[15] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4334–4343, 2024. 4

[16] Paul Debevec, Erik Reinhard, Greg Ward, and Sumanta Pattanaik. High dynamic range imaging. In *ACM SIGGRAPH 2004 Course Notes*, pages 14–es. 2004. 2, 3

[17] Piotr Didyk, Rafal Mantiuk, Matthias Hein, and Hans-Peter Seidel. Enhancement of bright video features for hdr displays. In *Computer Graphics Forum*, pages 1265–1274. Wiley Online Library, 2008. 3

[18] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021. 2

[19] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 257–266, 2002. 6, 8

[20] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017. 2, 3, 4, 6, 7, 8

[21] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177–1, 2017. 2, 3

[22] Mark D Fairchild. The hdr photographic survey. In *Color and imaging conference*, pages 233–238. Society of Imaging Science and Technology, 2007. 5

[23] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)*, 36(6):1–14, 2017. 5

[24] Abhishek Goswami, Aru Ranjan Singh, Francesco Banterle, Kurt Debattista, and Thomas Bashford-Rogers. Semantic aware diffusion inverse tone mapping. *arXiv preprint arXiv:2405.15468*, 2024. 3

[25] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 3

[26] Param Hanji, Rafal Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. Comparison of single image hdr reconstruction methods—the caveats of quality assessment. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–8, 2022. 6

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6, 8

[28] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 2, 3, 5, 8

[29] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 2

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[31] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[32] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[33] Lingtong Kong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu, and Jinwei Chen. Safnet: Selective alignment fusion network for efficient hdr imaging. *arXiv preprint arXiv:2407.16308*, 2024. 2, 3

[34] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018. 2, 3

[35] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019. 2

[36] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4758–4766, 2018. 6, 8

[37] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1651–1660, 2020. 2, 3, 5, 6, 7, 8

[38] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *European Conference on Computer Vision*, pages 344–360. Springer, 2022. 2, 3, 8

[39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 4

[40] Rafal K Mantiuk, Dounia Hammou, and Param Hanji. Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625*, 2023. 6

[41] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, pages 37–49. Wiley Online Library, 2018. 2, 3, 6, 7, 8

[42] Belen Masia, Sandra Agustin, Roland W Fleming, Olga Sorkine, and Diego Gutierrez. Evaluation of reverse tone mapping through varying exposure conditions. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–8. 2009. 3

[43] Belen Masia, Ana Serrano, and Diego Gutierrez. Dynamic range expansion based on image statistics. *Multimedia Tools and Applications*, 76:631–648, 2017. 3

[44] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 382–390. IEEE, 2007. 2, 4

[45] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16190–16199, 2022. 5

[46] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[47] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021. 2

[48] Karen Panetta, Landry Kezebou, Victor Oludare, Sos Agaian, and Zehua Xia. Tmo-net: A parameter-free tone mapping operator using generative adversarial network, and performance benchmarking on large scale hdr dataset. *IEEE Access*, 9:39500–39517, 2021. 5

[49] K Ram Prabhakar, Gowtham Senthil, Susmit Agrawal, R Venkatesh Babu, and Rama Krishna Sai S Gorthi. Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4875–4885, 2021. 2

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[52] Erik Reinhard and Kate Devlin. Dynamic range reduction inspired by photoreceptor physiology. *IEEE transactions on visualization and computer graphics*, 11(1):13–24, 2005. 6, 8

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans,

et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[55] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Transactions on Graphics (TOG)*, 39(4):80–1, 2020. 2, 3, 6, 7, 8

[56] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203–1, 2012. 2, 3

[57] Jou Won Song, Ye-In Park, Kyeongbo Kong, Jaeho Kwak, and Suk-Ju Kang. Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In *European Conference on Computer Vision*, pages 288–304. Springer, 2022. 2, 3, 8

[58] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023. 5, 6

[59] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14214–14223, 2023. 2

[60] Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, and Dominique Ginhac. Alignment-free hdr deghosting with semantics consistent transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12790–12799. IEEE, 2023. 2, 3, 5, 8

[61] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[62] Chao Wang, Ana Serrano, Xingang Pan, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt, and Thomas Leimkühler. Glowgan: Unsupervised learning of hdr images from ldr images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10509–10519, 2023. 2, 3, 6, 7, 8

[63] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018. 2, 3, 8

[64] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 4

[65] Pengfei Xiong and Yu Chen. Hierarchical fusion for practical ghost-free high dynamic range imaging. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4025–4033, 2021. 2

[66] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[67] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 2, 3

[68] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29: 4308–4322, 2020.

[69] Qingsen Yan, Song Zhang, Weiye Chen, Hao Tang, Yu Zhu, Jinqiu Sun, Luc Van Gool, and Yanning Zhang. Smae: Few-shot learning for hdr deghosting with saturation-aware masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5775–5784, 2023.

[70] Qian Ye, Jun Xiao, Kin-man Lam, and Takayuki Okatani. Progressive and selective fusion network for high dynamic range imaging. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5290–5297, 2021. 2, 3

[71] Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou, and Chunxia Xiao. Luminance attentive networks for hdr image and panorama reconstruction. In *Computer Graphics Forum*, pages 181–192. Wiley Online Library, 2021. 2, 3

[72] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2

[73] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[74] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2

[75] Ning Zhang, Yuyao Ye, Yang Zhao, and Ronggang Wang. Revisiting the stack-based inverse tone mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9162–9171, 2023. 2, 3

[76] Xuaner Zhang, Kevin Matzen, Vivien Nguyen, Dillon Yao, You Zhang, and Ren Ng. Synthetic defocus and look-ahead autofocus for casual videography. *ACM Transactions on Graphics (TOG)*, 38(4):1–16, 2019. 5

[77] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17907–17917, 2022. 2

[78] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019.

[79] Xinge Zhu, Zhichao Yin, Jianping Shi, Hongsheng Li, and Dahua Lin. Generative adversarial frontal view to bird view synthesis. In *2018 International conference on 3D Vision (3DV)*, pages 454–463. IEEE, 2018. 2