Diffusion-Based HDR Reconstruction from Mosaiced Exposure Images

Seeha Lee[a], Dongyoung Choi[b] and Min H. Kim[c]
School of Computing, KAIST, Daejeon, South Korea
shlee2@vclab.kaist.ac.kr, dychoi@vclab.kaist.ac.kr, minhkim@vclab.kaist.ac.kr

Keywords: Snapshot HDR imaging, diffusion-based image restoration

Abstract: Snapshot-based HDR imaging from Bayer-patterned multi-exposure inputs has gained significant attention with recent advancements in HDR imaging technology. Learning-based approaches have enabled the reconstruction of HDR images from extremely sparse multi-exposure measurements captured on a single Bayer-patterned sensor. However, existing learning-based methods predominantly rely on tone-mapped representations due to the inherent challenges of direct supervision in the HDR radiance domain. This tone-mapping-based approach suffers from critical limitations, including amplified noise and structural distortions in the reconstructed HDR images. The fundamental challenge arises from the high dynamic range of HDR radiance values, which exhibit a sparse and uneven distribution in floating-point space, making gradient-based optimization unstable. To address these issues, we propose a novel diffusion-based HDR reconstruction framework that operates directly in a split HDR radiance domain while preserving the linearity of the original HDR radiance values. By leveraging the generative power of diffusion models, our approach effectively learns the structural and radiometric characteristics of HDR images, leading to superior detail preservation, reduced noise artifacts, and enhanced reconstruction fidelity. Experiments demonstrate that our method outperforms state-of-the-art techniques in both qualitative and quantitative evaluations.

1 INTRODUCTION

High-dynamic-range (HDR) imaging is essential in computational photography, enabling accurate scene representations by capturing a wide range of radiance values. Unlike standard low-dynamic-range (LDR) images, which suffer from sensor limitations, HDR images store floating-point linear radiance values, making them critical for applications such as autonomous driving (Kocdemir et al., 2022; Seger, 2016), medical imaging (Kumar et al., 2023; Ramponi et al., 2016), and environmental monitoring (Suh et al., 2018; Jacobs, 2007). However, conventional sensors struggle to capture this dynamic range in a single shot—short exposures introduce noise, while long exposures lead to saturation and motion blur. To overcome these limitations, multi-exposure HDR reconstruction techniques (Debevec and Malik, 1997) merge LDR images captured at different exposure levels to recover lost radiance information.

Recently, snapshot-based HDR imaging has gained attention, leveraging Bayer-patterned multi-exposure inputs to capture multiple exposures in a single shot using specialized sensor designs such as quad-Bayer and multi-exposure CFA layouts (Kim and Kim, 2023; Suda et al., 2020; Jiang et al., 2021). This approach eliminates motion artifacts in sequential HDR methods while reducing capture la-

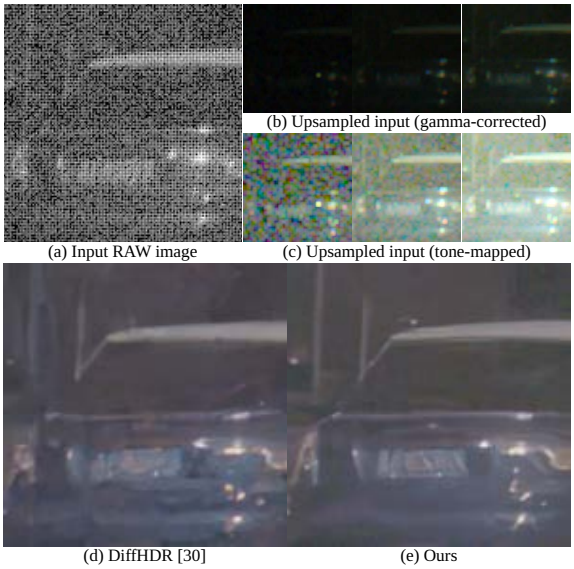


Figure 1: (a) Input quad-Bayer patterned RAW images with varying exposure times. (b) & (c) Three multi-exposed input images. (d) & (e) Results from DiffHDR and our method, respectively. The results demonstrate that our method reconstructs fine details.

[a] https://orcid.org/0009-0008-7382-5878
[b] https://orcid.org/0000-0003-1896-4038
[c] https://orcid.org/0000-0002-5078-4005

tency. However, reconstructing HDR images from sparse multi-exposure measurements remains challenging due to spatial resolution loss, increased sensor noise, and complex exposure relationships. Deep learning has enabled high-quality HDR reconstruction by learning to infer missing radiance values, yet still faceing challenges in optimization stability, noise robustness, and accurate radiance estimation.

A key challenge in learning-based HDR reconstruction is the extremely wide dynamic range of HDR radiance values, which makes gradient-based optimization unstable. To address this, prior works rely on tone-mapped supervision, such as the μ-law function (Kalantari et al., 2017), which logarithmically compresses radiance values. While this improves gradient flow, it alters pixel distributions, distorts contrast, and introduces floating-point precision errors. Training in the non-linear tone-mapped space also biases networks toward the compressed representation, leading to inconsistencies when converting outputs back to linear form, ultimately degrading HDR reconstruction accuracy.

Beyond tone-mapping issues, camera sensor noise variation further complicates HDR reconstruction. Noise scales with light intensity, making darker regions significantly noisier than brighter ones. This imbalance makes denoising more challenging in HDR domains compared to LDR domains to pose additional challenges, as most existing methods are trained on well-lit scenes with minimal noise, limiting their robustness in real-world conditions where exposure-dependent noise is prevalent.

To overcome these challenges, we propose a novel diffusion-based HDR reconstruction framework that directly operates in a split linear HDR radiance domain, eliminating tone-mapping artifacts. Following recent snapshot HDR imaging approaches (Kim and Kim, 2023; Suda et al., 2020; Jiang et al., 2021), our method uses raw quad-Bayer LDR images, which inherently capture multi-exposure information but suffer from spatial resolution loss and sensor noise.

To stabilize optimization in the linear domain, we introduce a range-splitting algorithm that partitions the HDR radiance domain into bounded intensity segments. This approach ensures stable training while effectively handling extreme intensity variations. Additionally, range-splitting applied to LDR inputs helps control sensor noise within each segment, enabling more robust feature extraction.

To further enhance reconstruction quality, we incorporate an LDR refinement network that denoises and processes segmented inputs, preventing noise amplification in low-exposure images. Instead of aligning all exposures to the shortest exposure (which has the highest noise), we leverage a self-attention module to extract reliable information adaptively from higher exposures with lower noise. These refined features condition our diffusion-based HDR network, enabling high-quality HDR generation in the linear radiance space.

In summary, our approach introduces a diffusion-based HDR reconstruction framework that operates in the linear HDR radiance domain, avoiding tone-mapping artifacts while achieving stable training through range-splitting. By incorporating an LDR refinement network and self-attention-based feature extraction, our method effectively suppresses sensor noise and enhances HDR quality. Extensive experiments demonstrate that our approach outperforms state-of-the-art HDR reconstruction methods in both synthetic and real-world datasets.

2 RELATED WORK

Multi-shot HDR reconstruction. Traditional HDR imaging merges multiple LDR images captured at different exposure levels into a single HDR image. (Kalantari et al., 2017) introduced a widely used dataset that aligns LDR images to a middle exposure reference. While effective for static scenes, this dataset does not account for real-world motion blur in longer exposures, leading to ghosting artifacts. Moreover, aligning HDR images to the middle exposure is suboptimal, as real-world scenes often contain motion blur in mid-exposure images. Aligning to the shortest exposure, which has the least motion blur, is a more practical approach. Additionally, since the dataset consists of well-lit scenes with minimal noise, it lacks sensor noise variability encountered in low-light environments, limiting model generalizability. To address these limitations, (Chi et al., 2023) introduced synthetic noise into the Kalantari dataset and trained a Swin-Transformer (Liu et al., 2021b) for denoising HDR reconstruction. However, it does not model real-world motion blur, reducing practical applicability.

Snapshot HDR reconstruction with quad-Bayer images. Recent advances in HDR imaging leverage quad-Bayer imaging sensors, which contain multiple exposures in a single shot. (Akyüz et al., 2020) employ a two-step method: a neural network first generates multiple LDR images from the raw quad-Bayer data, which are then merged using a conventional HDR reconstruction method. Although this method is promising for static scenes, it lacks an algorithm to mitigate ghosting artifacts in dynamic scenes. More recently, (Kim and Kim, 2023), (Suda et al., 2020), and (Jiang et al., 2021) proposed approaches that syn-

thesize realistic quad-Bayer LDR training data from the ALEXA HDR video dataset (Froehlich et al., 2014). Their transformer-based HDR reconstruction framework, aligned to the shortest exposure, improved robustness against motion artifacts. However, noise sensitivity in darker regions remains a significant challenge. Our approach builds on these approaches by employing diffusion models to directly model the distribution of linear HDR images, enhancing robustness in challenging lighting conditions.

Deep Learning-Based HDR reconstruction. Since (Debevec and Malik, 1997) introduced multi-exposure HDR imaging, research has shifted from simple dynamic range recovery to addressing data loss caused by motion blur, sensor noise, and misalignment. While short exposures retain highlight details but suffer from noise, long exposures recover shadows but introduce motion blur. Deep learning has significantly advanced HDR reconstruction by mitigating these issues through feature alignment and adaptive fusion. (Kalantari et al., 2017) pioneered deep learning-based HDR reconstruction using optical flow (Liu et al., 2009) for LDR alignment, followed by CNN-based merging. Later approaches refined this pipeline with feature-level alignment (Wu et al., 2018; Yan et al., 2020) and attention mechanisms for more precise fusion (Yan et al., 2019; Liu et al., 2021a). (Liu et al., 2022) further improved HDR reconstruction by integrating vision transformers (Dosovitskiy et al., 2021), effectively capturing both local and global dependencies. Despite these advances, existing deep learning methods struggle with extreme exposure variations and noise, particularly in low-light conditions. Instead of relying on deterministic alignment, our approach leverages diffusion-based generative modeling, enabling robust HDR reconstruction without explicit alignment constraints.

Generative Model-Based HDR reconstruction. Generative models have recently been explored for HDR reconstruction by learning the underlying HDR data distribution. (Niu et al., 2021) used a GAN-based model with a multi-scale generator and convolutional discriminator to enhance HDR image fidelity. (Yan et al., 2023) introduced diffusion models for HDR reconstruction, conditioning on LDR inputs. However, their approach trains in a tone-mapped HDR space, producing non-linear HDR estimates that require inversion, making the model highly dependent on the tone-mapping used during training. This dependency limits generalization, particularly in scenes with diverse lighting conditions.

In contrast, we train our diffusion model directly in linear HDR space, ensuring physically accurate radiance reconstruction. Our range-splitting strategy further stabilizes training by segmenting HDR radiance into bounded intensity ranges, preventing bias toward low-intensity values. This allows our model to capture fine details across the full exposure spectrum, outperforming existing generative approaches in HDR reconstruction.

3 METHOD

3.1 Preliminaries

Diffusion Model. The diffusion model is a generative model capable of transforming Gaussian noise into a specific data distribution. This transformation begins by defining the forward process, which iteratively adds small Gaussian noise to a clean image $\mathbf{x}_0$, until it converges to a standard normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I}) \sim \mathbf{x}_T$. Then, the diffusion model trains a noise estimation model that learns to reverse this process, referred to as the backward process. This process can be guided by a certain condition, enabling the diffusion model to generate samples based on that condition. Specifically, the forward process samples $\mathbf{x}_t$, the latent variable at timestep $t \in [1,\dots,T]$, from $\mathbf{x}_0$:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon_t, \tag{1}$$

where $\alpha_{1:T} \in (0,1)$ is a constant noise schedule that controls the noise level of each forward step, $\bar{\alpha}_t = \prod_{k=1}^{t}\alpha_k$, and $\varepsilon_t \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. The backward process of this procedure can be modeled by a neural network:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\mathbf{y},\mathbf{x}_t,t)\right) + \sqrt{1-\bar{\alpha}_t}\varepsilon, \tag{2}$$

where $\mathbf{y}$ is the input condition, and $\varepsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. Although ε plays a part in achieving sample diversity, it can hurt the consistency when the model is trained for image reconstruction. This can be resolved by using DDIM (Song et al., 2021) sampling instead:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\tilde{\mathbf{x}}_{0,t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\varepsilon_\theta(\mathbf{y},\mathbf{x}_t,t) + \sigma_t^2\varepsilon, \tag{3}$$

where σ_t controls how stochastic the sampling process is, and by setting it to 0, the sampling process becomes deterministic, more suitable for image reconstruction. Note that $\tilde{\mathbf{x}}_{0,t}$ is a temporal version of $\mathbf{x}_0$ obtained by replacing ε_t to ε_θ in Eq. (1):

$$\tilde{\mathbf{x}}_{0,t} = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_\theta(\mathbf{y},\mathbf{x}_t,t)}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

Noise Model. Camera sensor noise in raw Bayer images consists of shot noise, which arises from the statistics of photon arrivals at the sensor, and read noise, which stems from inaccuracies in the electronic circuits while reading the signal. Shot noise depends

on the captured signal, and it can be modeled as a Poisson random variable with the mean corresponding to the light intensity. On the other hand, read noise is independent of the signal and can be modeled using a Gaussian distribution with 0 mean, and a fixed variance. We can simulate these noises with a Gaussian distribution:

$$\mathbf{I}_{\text{noisy}} \sim \mathcal{N}(\mathbf{I}, \sigma_s^2 \mathbf{I} + \sigma_r^2), \tag{5}$$

$$\log(\sigma_s^2) \sim \mathcal{U}(\log(0.0012), \log(0.0048)), \tag{6}$$

$$\log(\sigma_r^2) \sim \mathcal{N}(1.869 \log(\sigma_s^2) + 0.3276, 0.3^2), \tag{7}$$

where $\mathbf{I}$ is the clean image in linear space, and σ_s and σ_r are standard deviation for the shot and read noise, respectively. We use the σ_s, σ_r values from a Xiaomi 10S smartphone (Zhang et al.,), and sample new noise levels in every training epoch.

3.2 Range Splitting Algorithm

Splitting Range. Our range-splitting algorithm divides a linear image $\mathbf{I}$ with a range of $[0,1]$ into segments $\mathbf{I}^{(1:S)}$. Here, S, the split level, represents the number of segments resulting from the split, and $r_{1:S}$ indicates the boundaries of the segments. Note that $r_1 = 0$ and $r_S = 1$. The s-th segment, $\mathbf{I}^{(s)}$, is defined by:

$$\mathbf{I}^{(s)} = \frac{\text{clamp}(\mathbf{I}, r_s, r_{s+1}) - r_s}{r_{s+1} - r_s}. \tag{8}$$

This procedure is applied to each RGB channel separately, resulting in an array of segments with $3S$ channels.

Combining Ranges. The range-combining process integrates the segments $\mathbf{I}^{(1:S)}$ back into the original linear image $\mathbf{I}$. The ground truth segments consist of clamped pixel values of exactly 0 or 1. However, in the segments predicted by the neural network, these pixel values are close to 0 or 1, but are not exactly equal. This discrepancy introduces noise into the combined image, especially when the combining process relies on a naive inverse formula of the range-splitting, thereby diminishing the quality of the predicted result. Therefore, we employ a strategy that eliminates pixel values belonging to other segments by applying masks to each segment. These masks are predicted using a neural network, and the ground truth masks $\mathbf{M}^{(1:S)}$ can be easily derived during the range-splitting process as:

$$\mathbf{M}^{(s)} = \begin{cases} 1, & \text{where } r_s \le \mathbf{I} < r_{s+1} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Summing up, the range-combining process is given by:

$$\mathbf{I} = \sum_{s \in [1:S]} \mathbf{M}_\phi^{(s)} \cdot \left(\mathbf{I}^{(s)} (r_{s+1} - r_s) + r_s \right), \tag{10}$$

Impact of our range-splitting. Applying range-splitting to the target HDR images divides the extensive dynamic range into smaller segments. This method enables optimization directly in linear space, eliminating the need for non-linear tone-mapping functions during training. In the LDR domain, range-splitting offers two benefits. First, it limits the noise level within each segment, making it easier to manage varying degrees of sensor noise. Second, it provides a more effective representation of both dark and bright regions compared to gamma correction, allowing the neural network to process them more efficiently. Together, these advantages enhance the overall quality of HDR reconstruction.

3.3 Splitting HDR Diffusion

Our model is a conditional diffusion model designed to produce linear HDR samples from noisy quad-Bayer patterned LDR images as input. By applying the proposed range-splitting algorithm to the input and using it to guide the diffusion process, our model has the capability to generate linear HDR samples in segments, which are then processed using the range-combining algorithm to obtain the linear HDR image. Our pipeline consists of three neural networks: LRNet, HDRNet, and MaskNet. The LRNet and HDRNet share the same U-Net backbone with different hyperparameters, while MaskNet is a lightweight CNN. The detailed architecture of the networks is in the supplemental document. Moreover, we train our model by cropping random patches from the full image to reduce computational cost and accommodate inputs of various sizes. An overview of our model is shown in Figure 3.

Preprocessing LDR Input. We convert the quad-Bayer patterned LDR image $\mathbf{L}_{\text{tetra}}$, into three separate images with different exposures. This process begins by subsampling the input LDR image of size $[H \times W]$ into RGB images that have different exposures of size $[3 \times H/4 \times W/4]$. These images are then upsampled using bilinear interpolation to recover the original image size, resulting in $\mathbf{L}_{\text{sht}}, \mathbf{L}_{\text{mid}}, \mathbf{L}_{\text{lng}}$, where each having a size of $[3 \times H \times W]$. The proposed range-splitting algorithm defined in Equation (8) is applied to these processed inputs, creating LDR segments $\mathbf{L}_{\text{sht}}^{(1:S)}, \mathbf{L}_{\text{mid}}^{(1:S)}, \mathbf{L}_{\text{lng}}^{(1:S)}$ with sizes of $[3S \times H \times W]$, respectively, where S is the total number of split segments.

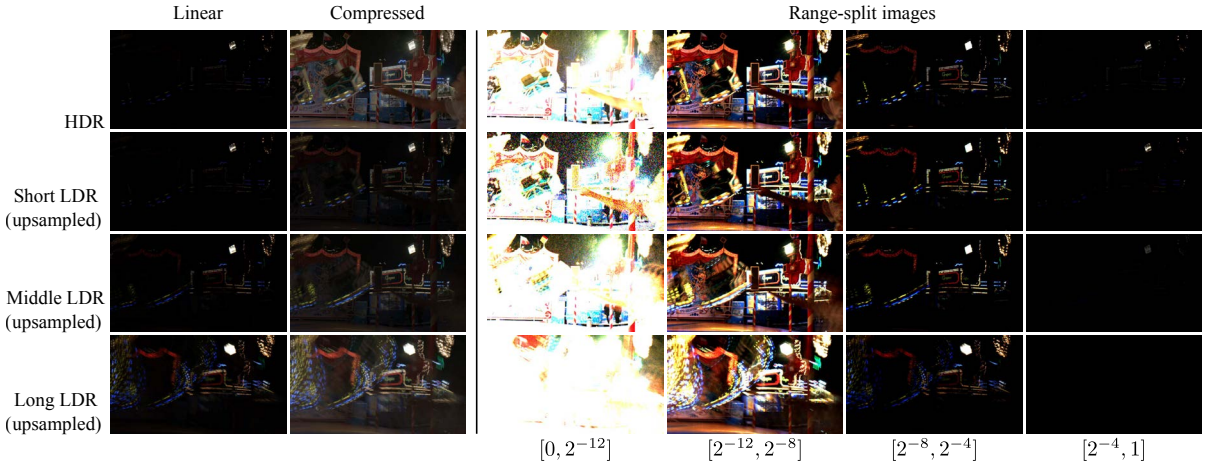


Figure 2: The proposed range-splitting algorithm is illustrated here. The first column displays the linear images of the corresponding inputs. The second column presents a *tone-mapped* image with a factor of 5k for the HDR image and *gamma-corrected* image for the LDR images. The third to sixth column shows the range-split image with $r_{1:S} = \{0, 2^{-12}, 2^{-8}, 2^{-4}, 1\}$.

Extracting Refined LDR Features. To effectively utilize the LDR segments with data loss as a condition for the diffusion model, we process them through an LDR refinement network (LRNet) to obtain refined features. Each segment is processed in a distinct pipeline, where each network receives $\mathbf{L}^{(i)}_{\text{sht}}, \mathbf{L}^{(i)}_{\text{mid}}, \mathbf{L}^{(i)}_{\text{lng}}$ as input and is trained produce $\hat{\mathbf{H}}^{(i)}_{\text{ldr}}$, the prediction of their HDR counterpart $\mathbf{H}^{(i)}_0$. While conventional methods align features to a single reference frame, which is the short exposure input in our case, this is not appropriate for our framework since the short exposure image suffers from severe data loss from sensor noise, the quad-Bayer pattern, and quantization. Therefore, we incorporate a self-attention layer to learn the relationship between the different exposures, addressing data loss in the reference frame. As shown in Figure 3a, we used a U-Net based autoencoder with depth d, where the final output of the network is a set of features $\mathbf{c}^{(i)}_{1:d}$ from each level in the encoding path. LRNet is optimized using a reconstruction loss for each split segment:

$$\mathcal{L}_{\text{ldr}} = \frac{1}{S}\sum_{i=1}^{S} \|\mathbf{H}^{(i)}_0 - \hat{\mathbf{H}}^{(i)}_{\text{ldr}}\|. \tag{11}$$

Generating Linear HDR Samples. Our conditional diffusion model produces linear HDR segments $\hat{\mathbf{H}}^{(1:S)}_0$, conditioned on LDR features $\mathbf{c}_{1:d}$. Adopting the strategy from (Li et al., 2022), we generate the residual $\mathbf{R}^{(1:S)}_0$ between the HDR image and the upsampled short LDR image to accelerate model convergence. Moreover, inspired by (Zhang et al., 2023), we implement a hierarchical conditioning method, where the conditional features are applied to the decoder path of the diffusion U-Net, HDRNet, at each level, providing a stronger condition for the diffusion process. Our model is optimized with the following loss function:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{noise}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}}, \tag{12}$$

where $\mathcal{L}_{\text{noise}}$ is the conventional diffusion loss, $\mathcal{L}_{\text{img}}$ is an image space loss, and λ_{img} is a weight for the image space loss. Similar to (Yang et al., 2023), our image space loss uses the temporal prediction $\tilde{\mathbf{R}}^{(1:S)}_{0,t}$ defined in Equation (4), minimizing the color distortion in the sampled result:

$$\mathcal{L}_{\text{img}} = \mathbb{E}_t\left[\|\mathbf{R}^{(1:S)}_0 - \tilde{\mathbf{R}}^{(1:S)}_{0,t}\|\right]. \tag{13}$$

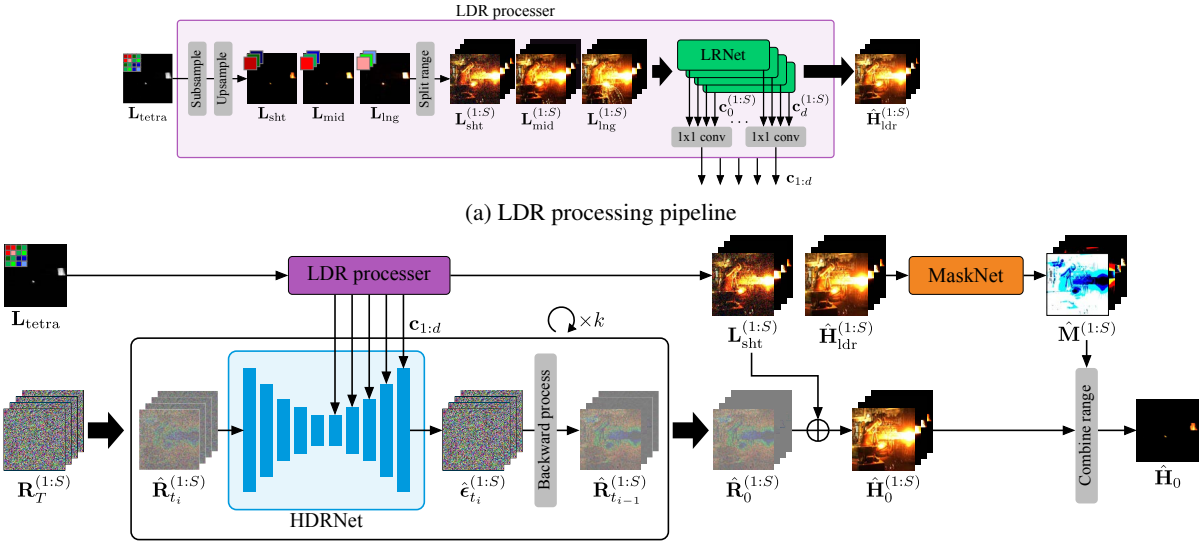
For effective sampling, we employ the DDIM method (Song et al., 2021), reducing the sampling steps to $\{t_k, \ldots, t_1\}$. The entire pipeline is illustrated in Figure 3b.

Mask Prediction. Our mask model, MaskNet, aims to generate a probability matrix $\mathbf{p}^{(1:S)}$ corresponding to the reference mask $\mathbf{M}^{(1:S)}$ of the split HDR image. The predicted mask $\hat{\mathbf{M}}^{(1:S)}$ can be obtained by selecting the index with the highest probability:

$$\hat{\mathbf{M}}^{(s)} = \begin{cases} 1 & \text{if } \underset{i\in[1:S]}{\text{argmax}}\ \mathbf{p}^{(i)} = s \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The mask model predicts masks using the $\hat{\mathbf{H}}^{(i)}_{\text{ldr}}$ as input, which is optimized by cross-entropy loss between $\mathbf{p}^{(1:S)}$ and $\mathbf{M}^{(1:S)}$:

$$\mathcal{L}_{\text{mask}} = -\frac{1}{S}\sum_{s=1}^{S} \log \frac{\exp(\mathbf{p}^{(s)}) \cdot \mathbf{M}^{(s)}}{\sum_{i=1}^{S}\exp(\mathbf{p}^{(i)})}. \tag{15}$$



(a) LDR processing pipeline

(b) HDR diffusion pipeline

Figure 3: Overview of the architecture of our model. Figure (a) illustrates the LDR processing pipeline, where the input quad-Bayer patterned image is processed to generate refined LDR features and a segment-wise mask. Figure (b) depicts the HDR diffusion pipeline, which utilizes the refined features as a condition to train the diffusion model. The model is trained to predict the range-split segments of the residual $\mathbf{R_0}$ between the HDR image $\mathbf{H}_0$ and the short-exposed LDR image $\mathbf{L}_{\text{sht}}$, subsampled from the quad-Bayer input. The desired HDR sample can be obtained by applying the range-combining algorithm to the predicted residual segments.

Predicted mask Reference mask

Figure 4: Visualization of predicted and reference masks.

We visualize the predicted mask from the mask model in Figure 4.

Patch-based Sampling. Our patch-based approach generates borders between the patches. Therefore, we sample the entire image using overlapping patches, similar to WeatherDiffusion (Özdenizci and Legenstein, 2023). At each backward step t_i, the predicted noise $\varepsilon_\theta(\mathbf{c}_{1:d}, \mathbf{R}_{t_i}^{(1:S)}, t_i)$ of the patches is merged, where the overlapped regions are divided by the number of overlaps. This strategy effectively smooths out the predicted sample at each step, resulting in an image without border artifacts.

4 RESULT

4.1 Experiment Setup

Synthetic Multi-Exposure Experiment. We create quad-Bayer patterned LDR inputs based on (Kim and Kim, 2023). Using the ALEXA HDR video dataset (Froehlich et al., 2014), we add 1, 4, and 16 consecutive frames to produce varying exposures with a difference of two photographic stops between each exposure level. By applying the quad-Bayer patterned array, we can obtain a single linear image with varying exposures within the HDR range. This procedure is followed by dynamic range clipping, the addition of synthetic noise, and quantization to create an LDR version of this image. We utilize the noise model introduced in Section 3.1 and quantize the image to 12 bits, a standard achievable by camera sensors in smartphones. For the reference HDR image, we use the frame aligned with the shortest exposure. We obtain 320 images from 22 different scenes for training and 53 images from 4 different scenes for testing.

Real Multi-Exposure Experiment. Tetra-binning camera sensors are commonly found in modern smartphones; however, they only provide access to processed Bayer images, not the raw quad-Bayer data. Therefore, we create a quad-Bayer patterned input using a real dataset captured at night with the main camera of a Xiaomi 10S smartphone provided by (Zhang et al.,). Each scene in this dataset comprises five 10-bit images captured through exposure bracketing, with a 2-stop difference between consecutive exposures — aligning with the configuration of our synthetic dataset. We subsample the first three images

Figure 5: Qualitative comparison on synthetic HDR data. Our method reconstructs sharper details and preserves textures more effectively, particularly in darker regions. More images are in Section 4 of the supplemental document.

into a single quad-Bayer image, effectively generating input data with actual sensor noise. Note that there is no ground-truth image for this real-world test dataset.

Implementation Details. We trained our model for 500k iterations, using a batch size of 16. In each iteration, we crop random patches of size 128×128 from every image. We utilize the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 0.02$, and a fixed learning rate of 0.00005. We implemented our model using PyTorch Lightning, trained on two NVIDIA A100 GPUs. The diffusion model is sampled using 25 DDIM steps, and a 32-pixel overlap.

4.2 Comparison

We evaluate the performance of our proposed model against state-of-the-art techniques. We select representative methods from each major deep learning approach: AHDR (Yan et al., 2019) for attention-based methods, HDRGAN (Niu et al., 2021) for GAN-based methods, and DiffHDR (Yan et al., 2023) for diffusion-based methods. These methods all use ordinary LDR inputs, not quad-Bayer inputs. For transformer-based methods, regarding the vast number of works in this domain, we choose 3 papers; CA-ViT (Liu et al., 2022), SCTNet (Tel et al., 2023) and (Kim and Kim, 2023). It is important to note that except from (Kim and Kim, 2023) which aligns with our use of quad-Bayer inputs, all other baseline methods are designed for standard LDR inputs. We fix the synthetic noise level to the midpoint of the noise range:

$$\sigma_s^2 = 0.0024,\ \log(\sigma_r^2) = 1.869\log(\sigma_s^2) + 0.3276, \tag{16}$$

which is applied to the tetra-binning LDR input using Equation (5). Since the method by (Kim and Kim, 2023) uses raw quad-Bayer images as input, we provided $\mathbf{L}_{\text{tetra}}$ from our new dataset. For the remaining methods, which take three LDR images as input, we followed the preprocessing procedure described in Section 3.3, excluding the range-splitting step, and used the upsampled LDR stack — $\mathbf{L}_{\text{sht}}$, $\mathbf{L}_{\text{mid}}$, and $\mathbf{L}_{\text{lng}}$ — as input. Additionally, we modified the alignment process of each method to consider that our dataset uses the shortest exposure ($\mathbf{L}_{\text{sht}}$) as the alignment anchor, differing from the Kalantari dataset, which aligns to the middle exposure ($\mathbf{L}_{\text{mid}}$). All methods were trained until convergence using a μ (tone-mapping factor) value of 5000, which is a common setting in HDR imaging tasks.

Table 1 presents the quantitative results of the methods. We evaluate the accuracy of the results using 9 evaluation metrics. 4 commonly used metrics : PSNR, PSNR-μ (Kalantari et al., 2017), SSIM, SSIM-μ (Wang et al., 2004), 4 HDR-specific metrics : pu21-PSNR, pu21-MSSSIM, pu21-VSI (Azimi et al., 2021), HDR-VDP-2 (Mantiuk et al., 2011), and NoR-VDP (Banterle et al., 2020) a metric for no-reference situation. The subscript μ indicates that the metric is computed in the tone-mapped domain, where we set $\mu = 5000$ to match the experimental conditions used in the compared methods. Moreover, we set the peak value to $4,000$ for the pu21 metrics. Our proposed method achieves the best performance in all HDR-specific metrics, and in most of the common metrics except for SSIM-μ, where it ranks second with a score very close to the best result. Figure 5 illustrates the qualitative results. Our method preserves more detail, particularly in darker regions, compared to other methods. Refer to the Appendix for more figures from synthetic data.

Cross-Validation. To further evaluate the generalization ability of our method, we conduct qualitative cross-validation using unseen data from our real dataset, which are shown in Figure 6. Compared to other methods, our approach preserves fine details while avoiding blurry artifacts, demonstrating strong robustness to real-world degradation. These results show the effectiveness of our framework in handling

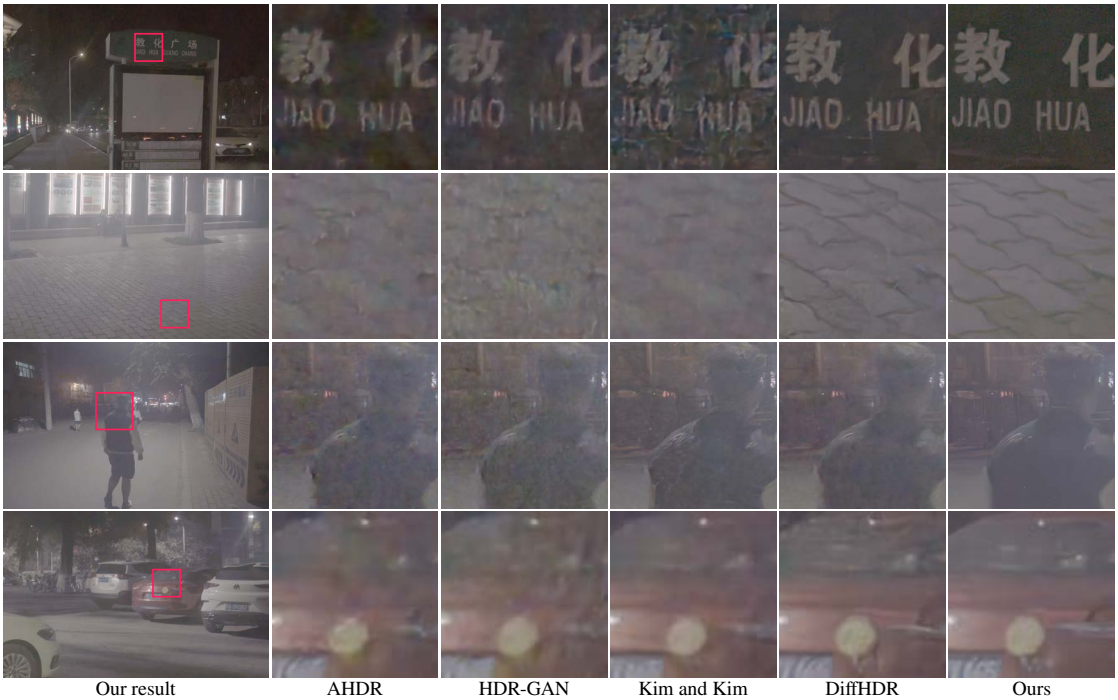


Figure 6: Qualitative comparison on real-world HDR data. Our framework reconstructs finer details and preserves textures more effectively than prior methods, demonstrating strong generalization to unseen data.

Figure 7: Additional results on real-world data. Compared to DiffHDR, our model generates sharper and cleaner HDR reconstructions, effectively handling real-world noise.

Table 1: Quantitative comparison on synthetic and real datasets (**bold red**: best, **bold blue**: second-best).

	Synthetic								Real
	PSNR	PSNR-μ	SSIM	SSIM-μ	pu21-PSNR	pu21-MSSSIM	pu21-VSI	HDR-VDP-2	NoR-VDP
AHDR	43.15	36.88	**0.9953**	**0.8938**	32.4701	0.9605	0.9947	59.08	58.02
HDRGAN	43.92	36.99	0.9932	0.8892	32.2663	0.9621	0.9951	59.32	59.03
CA-ViT	44.13	36.72	0.9946	0.8896	32.5615	0.9646	0.9954	58.87	58.12
SCTNet	44.48	36.90	0.9948	0.8910	32.6134	0.9653	0.9955	59.25	58.35
Kim and Kim	44.68	36.98	0.9948	0.8927	32.7485	0.9671	0.9960	60.57	59.06
DiffHDR	**45.62**	**37.12**	**0.9953**	0.8929	**33.0884**	**0.9678**	**0.9964**	**60.92**	**59.43**
Ours	**47.41**	**37.50**	**0.9956**	**0.8936**	**33.2015**	**0.9693**	**0.9969**	**61.54**	**60.73**

challenging noise patterns and unseen data. Additional results from the real dataset are provided in Figure 7. We compare our result with the results from DiffHDR (Yan et al., 2023), which performed best among the baseline methods in the synthetic dataset. Our model shows cleaner results, proving our model can effectively produce liner HDR images robust to noise even in unseen real-world data.

4.3 Ablation Study

We conduct an ablation study to assess the effectiveness of the proposed range-splitting algorithm. Using the same network architecture, we train the model with various supervision strategies: directly using linear HDR images as the reference, employing tone-mapped HDR images, and applying our range-splitting approach. For the linear and tone-mapped cases, we adhere to previous methods by concatenating the linear and gamma-corrected LDR images as input to the LDR processing pipeline. As shown in Table 2, the network struggles to converge when trained directly with linear HDR images. Moreover, our range-splitting method surpasses the tone-mapped approach, demonstrating that the proposed range-splitting algorithm effectively facilitates stable and precise optimization in linear space. Through extensive experiments, we empirically determined that using three internal boundaries (at values of 2^{-12}, 2^{-8}, and 2^{-4}) achieves an optimal trade-off between accuracy and computational complexity.

We also conduct an ablation study on the use of masks in the range-combining process. Using the final trained model, we compare utilizing the predicted mask, omitting the mask, and employing the reference mask obtained from the ground truth HDR images. As shown in Table 2, applying masks during range-combining significantly enhances the quality of the predicted samples by effectively masking out saturated pixels. The reference mask outperforms the predicted mask, underscoring the critical role of mask prediction in achieving accurate HDR reconstruction.

Furthermore, we conduct an ablation study to validate our proposed LDR feature alignment strategy. We compare the output of LRNet of our self-attention based alignment against two key baselines: a conventional attention-based alignment to the short-exposure reference image, and a variant with no feature alignment. Due to the significant computational cost of training the diffusion model, this ablation was performed by training the LDR reconstruction network (LRNet) alone, not training the full end-to-end pipeline. As shown in Table 3, the conventional alignment method underperforms our approach, highlighting the difficulty of aligning features to a degraded, noisy reference image. Our self-attention strategy, in contrast, can adaptively leverage the cleaner signal from misaligned longer exposures, resulting in a more robust feature alignment.

Table 2: Quantitative comparison of different training strategies. Our proposed range-splitting approach achieves the highest performance, with ground-truth masks yielding the best results.

Range-splitting	Strategy	PSNR ↑	PSNR-μ ↑	SSIM-μ ↑
$\{0,1\}$ (No splitting)	Linear	34.48	12.16	0.0884
	Tone-map	44.87	36.96	0.8929
$\{0,2^{-8},1\}$	No mask	37.05	12.67	0.095
	Predicted mask	43.52	29.70	0.6838
	GT. mask	43.54	29.77	0.6851
$\{0,2^{-10},2^{-6},1\}$	No mask	39.14	13.27	0.1119
	Predicted mask	44.81	35.61	0.8503
	GT. mask	44.84	35.87	0.8567
$\{\mathbf{0},\mathbf{2^{-12}},\mathbf{2^{-8}},\mathbf{2^{-4}},\mathbf{1}\}$	No mask	40.50	13.43	0.1273
	Predicted mask	47.41	37.40	0.8936
	GT. mask	**47.45**	**37.94**	**0.9087**

Table 3: Ablation study on ldr feature alingment. All the networks are trained for 200k isolated from the diffusion model.

Strategy	PSNR	PSNR-μ	SSIM-μ
No alignment	42.57	36.42	0.8862
Alignment to reference image	44.24	36.83	0.8875
Adaptive alignment with self-attention	**44.64**	**36.91**	**0.8903**

5 CONCLUSION

We have introduced a novel diffusion-based HDR reconstruction framework that operates directly in the

linear HDR radiance domain, overcoming the limitations of traditional tone-mapped supervision. Our approach leverages quad-Bayer patterned inputs and a range-splitting algorithm, which partitions both input LDR images and target HDR images into bounded intensity segments. This strategy stabilizes optimization in linear space, effectively mitigating the biases introduced by tone-mapping. Additionally, by limiting noise levels within each segment, our framework improves denoising efficiency, enhancing overall reconstruction quality.

By integrating range-splitting with a diffusion model, our method learns the distribution of linear HDR images, enabling it to recover lost details caused by sensor noise, motion blur, and the unique challenges of tetra-binning sensors. Experimental results demonstrate that surpasses state-of-the-art HDR reconstruction methods, performing robustly in both synthetic benchmarks and cross-validation on unseen real-world data. This highlights its potential to bridge the gap between synthetic training and real-world HDR imaging applications.

Despite its advantages, our approach has a limitation in the final range-combining process, where reconstruction quality depends on the accuracy of the predicted masks. In this work, we employed a simple CNN for mask prediction, which may constrain the achievable quality. Future research could explore more advanced range-combining strategies, potentially further improving the fidelity of the reconstructed HDR images.

Acknowledgements

Min H. Kim acknowledges the Samsung Research Funding & Incubation Center (SRFC-IT2402-02), the Korea NRF grant (RS-2024-00357548), the MSIT/IITP of Korea (RS-2022-00155620, RS-2024-00398830, RS-2024-00436680, and 2017-0-00072), and Microsoft Research Asia.

REFERENCES

Akyüz, A. O. et al. (2020). Deep joint deinterlacing and denoising for single shot dual-iso hdr reconstruction. *IEEE Transactions on Image Processing*, 29:7511–7524.

Azimi, M. et al. (2021). Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE.

Banterle, F., Artusi, A., Moreo, A., and Carrara, F. (2020). Nor-vdpnet: A no-reference high dynamic range quality metric trained on hdr-vdp 2. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 126–130. IEEE.

Chi, Y., Zhang, X., and Chan, S. H. (2023). Hdr imaging with spatially varying signal-to-noise ratios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5724–5734.

Debevec, P. E. and Malik, J. (1997). Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH*.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*.

Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., and Brendel, H. (2014). Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In *Digital photography X*, volume 9023, pages 279–288. SPIE.

Jacobs, A. (2007). High dynamic range imaging and its application in building research. *Advances in building energy research*, 1(1):177–202.

Jiang, Y., Choi, I., Jiang, J., and Gu, J. (2021). Hdr video reconstruction with tri-exposure quad-bayer sensors. *arXiv preprint arXiv:2103.10982*.

Kalantari, N. K., Ramamoorthi, R., et al. (2017). Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1.

Kim, J. and Kim, M. H. (2023). Joint demosaicing and deghosting of time-varying exposures for single-shot hdr imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12292–12301.

Kocdemir, I. H., Akyuz, A. O., Koz, A., Chalmers, A., Alatan, A., and Kalkan, S. (2022). Object detection for autonomous driving: high-dynamic range vs. low-dynamic range images. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE.

Kumar, G. A., Rahul, G., Preejith, S., and Sivaprakasam, M. (2023). Improving endoscopic image quality through the use of high dynamic range imaging-like method with real-time performance. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE.

Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., and Chen, Y. (2022). Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59.

Liu, C. et al. (2009). *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology.

Liu, Z., Lin, W., Li, X., Rao, Q., Jiang, T., Han, M., Fan, H., Sun, J., and Liu, S. (2021a). Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proc. CVPR*, pages 463–470.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierar-

chical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Liu, Z., Wang, Y., Zeng, B., and Liu, S. (2022). Ghost-free high dynamic range imaging with context-aware transformer. In *Proc. ECCV*, pages 344–360. Springer.

Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. (2011). Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14.

Niu, Y., Wu, J., Liu, W., Guo, W., and Lau, R. W. (2021). HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896.

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

Ramponi, G., Badano, A., Bonfiglio, S., Albani, L., and Guarnieri, G. (2016). An application of hdr in medical imaging. In *High Dynamic Range Video*, pages 499–518. Elsevier.

Seger, U. (2016). Hdr imaging in automotive applications. In *High Dynamic Range Video*, pages 477–498. Elsevier.

Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Suda, T., Tanaka, M., Monno, Y., and Okutomi, M. (2020). Deep snapshot hdr imaging using multi-exposure color filter array. In *Proceedings of the Asian Conference on Computer Vision*.

Suh, H. K., Hofstee, J. W., and Van Henten, E. J. (2018). Improved vegetation segmentation with ground shadow removal using an hdr camera. *Precision Agriculture*, 19:218–237.

Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., and Ginhac, D. (2023). Alignment-free hdr deghosting with semantics consistent transformer. In *Proc. ICCV*, pages 12790–12799. IEEE.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Wu, S., Xu, J., Tai, Y.-W., and Tang, C.-K. (2018). Deep high dynamic range imaging with large foreground motions. In *Proc. ECCV*, pages 117–132.

Yan, Q., Gong, D., Shi, Q., Hengel, A. v. d., Shen, C., Reid, I., and Zhang, Y. (2019). Attention-guided network for ghost-free high dynamic range imaging. In *Proc. CVPR*, pages 1751–1760.

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

Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., and Zhang, Y. (2020). Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322.

Yang, S., Hwang, H., and Ye, J. C. (2023). Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proc. ICCV*, pages 22873–22882.

Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Zhang, Z., Zhang, S., Wu, R., Yan, Z., and Zuo, W. Exposure bracketing is all you need for a high-quality image. In *The Thirteenth International Conference on Learning Representations*.

APPENDIX

A Detailed network architecture

We use a U-Net-based architecture with self-attention blocks for both LRNet and HDRNet. The backbone consists of five levels, each containing two residual blocks. To better capture complex details, we incorporate self-attention blocks at the fourth and fifth levels. Both the residual and self-attention blocks are derived from (Dhariwal and Nichol, 2021). The hyperparameters for LRNet and HDRNet are detailed in Table 4.

Table 4: Hyperparameters for LRNet and HDRNet.

Parameter	**LRNet**	**HDRNet**
Number of Levels	5	5
Residual Blocks per Level	2	2
Self-Attention Levels	4th, 5th	4th, 5th
Initial Channel Count	64	256
Final Activation Function	Tanh	None
Use Scale-Shift Norm in Residual Blocks	False	True

B Experimental details of comparison

We compare our method against AHDR (Yan et al., 2019), CA-ViT (Liu et al., 2022), SCTNet (Tel et al., 2023), (Kim and Kim, 2023), HDRGAN (Niu et al., 2021), and DiffHDR (Yan et al., 2023). For AHDR, (Kim and Kim, 2023), and HDRGAN, we follow the original papers' instructions, using the same network architectures and experimental settings. However, for DiffHDR, we adopt the same HDR diffusion U-Net architecture as our framework and train the model for an equal number of iterations to ensure a fair comparison.

C Additional results from synthetic dataset

Additional results from the synthetic dataset are provided in Figures 8. The results show that our method reconstructs sharper details than baseline methods, particularly in darker scenes. This improvement comes from our range-splitting process, which effectively divides the wide noise range and confines the noise level within each split segment, simplifying the denoising task.

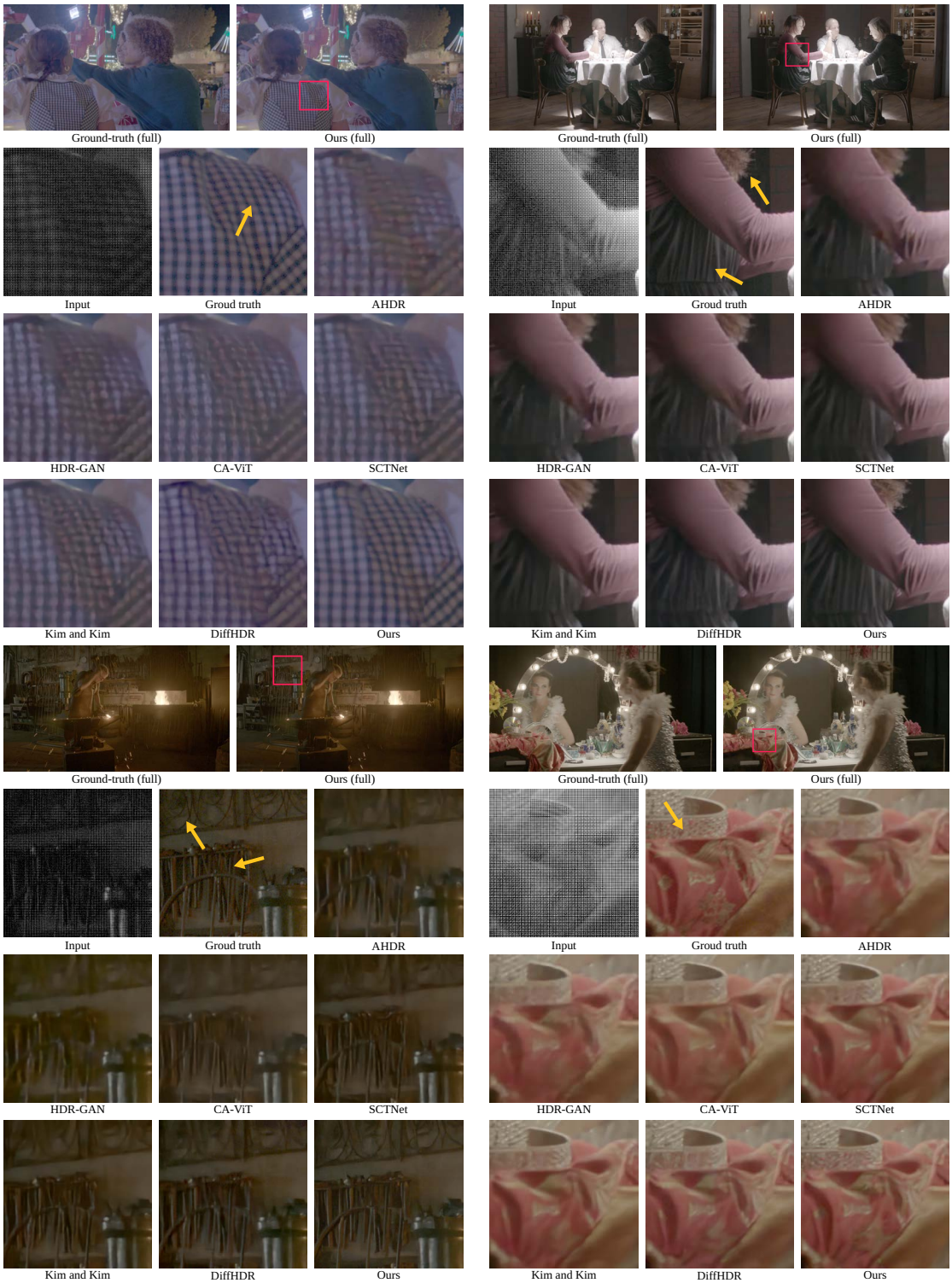


Figure 8: Additional results on synthetic data. All the images, including the input, are displayed using $\mu = 500,000$. Our method significantly outperforms baseline methods, producing sharper and more detailed reconstructions.