InstantHDR: Single-forward Gaussian Splatting for High Dynamic Range 3D Reconstruction

Dingqiang Ye*[1], Jiacong Xu*[1], Jianglu Ping[1], Yuxiang Guo[1], Chao Fan[2], and Vishal M. Patel†[,1]

[1] Johns Hopkins University, USA
[2] Shenzhen University, China
{dye6,jxu155,jping1,yguo87}@jhu.edu, chaofan996@szu.edu.cn,vpatel36@jhu.edu

Abstract. High dynamic range (HDR) novel view synthesis (NVS) aims to reconstruct HDR scenes from multi-exposure low dynamic range (LDR) images. Existing HDR pipelines heavily rely on known camera poses, well-initialized dense point clouds, and time-consuming per-scene optimization. Current feed-forward alternatives overlook the HDR problem by assuming exposure-invariant appearance. To bridge this gap, we propose InstantHDR, a feed-forward network that reconstructs 3D HDR scenes from uncalibrated multi-exposure LDR collections in a single forward pass. Specifically, we design a geometry-guided appearance modeling for multi-exposure fusion, and a meta-network for generalizable scene-specific tone mapping. Due to the lack of HDR scene data, we build a pre-training dataset, called HDR-Pretrain, for generalizable feed-forward HDR models, featuring 168 Blender-rendered scenes, diverse lighting types, and multiple camera response functions. Comprehensive experiments show that our InstantHDR delivers comparable synthesis performance to the state-of-the-art optimization-based HDR methods while enjoying $\sim$ 700$\times$ and $\sim$ 20$\times$ reconstruction speed improvement with our single-forward and post-optimization settings. All code, models, and datasets will be released after the review process.

Keywords: High Dynamic Range · 3D Gaussian Splatting · Novel-View Synthesis · Feed-Forward Models

1 Introduction

High dynamic range (HDR) novel view synthesis (NVS) aims to reconstruct HDR scenes from multi-view low dynamic range (LDR) images captured at varying exposures. Unlike typical low dynamic range (from 0 to 255) imaging, which often suffers from detail loss in extreme lighting and color distortion due to sensor limitations, HDR captures a broader spectrum of luminance (from 0 to $+\infty$). Through integrating advanced frameworks like NeRF [47] or 3D Gaussian Splatting [26] with a tone mapper to model the camera response function (CRF), HDR

* co-first authors; †corresponding author

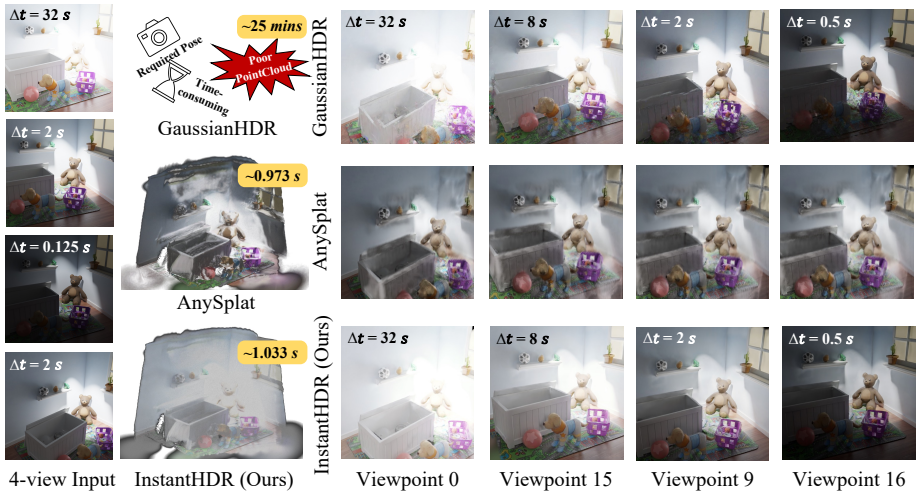


Fig. 1: Comparisons of reconstructed time (yellow boxes), scenes (left) and rendered views (right) between the GaussianHDR [35] (top), original AnySplat [18] (middle) and our InstantHDR (bottom). (i) GaussianHDR [35] spends expensive 25 *mins* and produces tearing artifacts, as its initial point clouds collapse under the sparse-view inputs. (ii) AnySplat [18] naively fuses multi-exposure inputs, causing ghosting artifacts and lacking exposure control. (iii) Our InstantHDR reconstructs 3D-consistent HDR scenes in few seconds and renders clean LDR images with controllable exposure time.

task enables the re-rendering of photo-realistic novel views with controllable exposure. Its ability to faithfully represent real-world light and shadow makes it indispensable for multiple applications such as autonomous driving [14,57,61,75], digital humans [13,36,86,87], and immersive image editing [30,37,56,80].

Existing HDR NVS methods [2,15,35] are predominantly optimization-based. However, this paradigm is inherently costly and fails to generalize, as it necessitates time-consuming per-scene optimization, precisely calibrated camera poses, and dense multi-view inputs for point cloud initialization. For example, as shown in Fig. 1, GaussianHDR [35] struggles to reconstruct scenes from only four LDR views, since exposure-induced appearance inconsistencies degrade the reliability of SfM-based point cloud initialization in sparse-view settings. Furthermore, their heavy computational overhead and strong data dependency limit practical deployment in real-time scenarios.

Recently, 3D feed-forward models [18,63] have revolutionized scene reconstruction by inferring geometry in seconds, achieving impressive speed and generalization than optimization-based methods. Integrating this paradigm into the HDR NVS tasks could boost model generalizability and inference speed. However, directly applying the original feed-forward models to HDR reconstruction may encounter the following issues. (a) **Exposure-induced Appearance Inconsistency**: Naive fusion leads to severe ghosting — as shown in Fig. 1. The

same white wall appears bright at $\Delta = 32s$ but nearly black at $\Delta = 0.125s$, causing visible artifacts in AnySplat [18]. (b) **Pixel-level Geometric Alignment**: Establishing accurate pixel-level geometric correspondences remains a non-trivial task under large brightness variations. (c) **Camera Response Functions Inconsistency**: In real world, different camera and software apply distinct color transformations (*e.g.*, AgX, Filmic), making it difficult to learn a unified tone mapping operator. (d) **HDR Data Scarcity**: Current publicly available HDR datasets [15,21] are insufficient (as shown in Tab. 1) to support the robust large-scale pre-training required for feed-forward models.

To address these challenges, we propose **InstantHDR**, a novel feed-forward 3D reconstruction framework for HDR novel view synthesis. InstantHDR comprises two key components. First, a *geometry-guided appearance modeling module* normalizes multi-exposure LDR inputs into a unified exposure space and utilizes geo-attention from the geometry encoder to fuse patch-level irradiance features. Fine-grained texture details are further recovered by incorporating Difference of Gaussians (DoG) high-frequency cues, lifting the representation to pixel-level irradiance before predicting HDR 3D Gaussians. Second, a *MetaNet* takes the predicted HDR Gaussians, LDR images, and exposure times as input to estimate scene-specific tonemapper parameters, enabling generalizable HDR-to-LDR rendering with controllable exposure time Δt. The entire pipeline operates in a single forward pass without per-scene optimization, significantly improving reconstruction speed over recent optimization-based methods. With an optional lightweight post-optimization step, InstantHDR achieves superior synthesis quality in sparse-view while maintaining competitive performance in dense-view.

Our contributions can be summarized as follows:

(i) We propose InstantHDR, the first feed-forward HDR novel view synthesis method. It features a geometry-guided appearance module for exposure-robust multi-view fusion and a meta-network for generalizable tone mapping, enabling 3D HDR reconstruction from uncalibrated multi-exposure LDRs in seconds.

(ii) We build HDR-Pretrain, a large-scale dataset of 168 synthetic indoor scenes to support feed-forward HDR pretraining. Experiments show InstantHDR achieves competitive quality while being $\sim$700$\times$ (single-forward) and $\sim$20$\times$ (post-optimization) faster than SoTA optimization-based HDR methods.

2 Related Works

High Dynamic Range Imaging. HDR imaging has traditionally been approached by merging multiple LDR exposures captured from a fixed viewpoint [45] or by recovering the camera response function from bracketed LDR sequences [6, 67]. While effective for static scenes, these methods suffer from ghosting artifacts when scene motion is present. Subsequent works [10,16,22,60,72] mitigate this by estimating optical flow to detect and compensate for motion prior to fusion. More recently, learning-based approaches leveraging CNNs [8, 27, 29, 40] and Transformers [3, 20, 41, 55, 78, 79] directly learn LDR-to-HDR mappings from data. Additionally, several methods reconstruct HDR from single LDR images using

handcrafted priors in an unsupervised or self-supervised manner [9,49,52,71,85]. However, all these methods operate in the 2D image domain and lack 3D scene understanding, making them unable to synthesize HDR views from novel view.

Gaussian Splatting. 3D Gaussian Splatting (3DGS) [25] represents scenes as collections of anisotropic Gaussian primitives, enabling real-time rendering through efficient rasterization—offering a significant speed advantage over NeRF-based volumetric ray-marching [15,21,46]. This efficiency has driven its adoption across diverse tasks including dynamic scenes [43,68,74], SLAM [24,44,70,81], inverse rendering [19,34,69], digital humans [12,31,38], 3D generation [33,58,77], and medical imaging [1,82]. Nevertheless, current HDR extensions [2,35] of 3DGS remain predominantly optimization-based, resulting in expensive per-scene reconstruction times. Our work aims to fill this research gap.

Feed-forward 3D Reconstruction. Recent advances pursue end-to-end 3D reconstruction directly from unposed images. Pioneering works such as DUSt3R [66] and MASt3R [32] replace traditional multi-stage pipelines with a unified model that jointly estimates depth and performs dense scene fusion. Subsequent approaches [39,48,59,62,63,65,73] extend this paradigm by cascading transformer blocks to simultaneously recover camera poses, point trajectories, and scene geometry in a single forward pass. A parallel line of research [4,11,17,54,64,76,84] targets novel view synthesis from unposed sparse-view images. These feed-forward models offer remarkable reconstruction speed, strong generalization to unseen scenes, and minimal data requirements compared to optimization-based pipelines. Their potential for HDR reconstruction remains largely unexplored. Our work explores this promising direction.

3 Method

Given uncalibrated multi-view LDR images captured at varying exposures, InstantHDR reconstructs HDR 3D Gaussians and renders novel views at any target exposure (Fig. 2). We first formalize the problem in Sec. 3.1, then detail the pipeline design in Sec. 3.2, including a Geo-guided Appearance Modeling module (Sec. 3.3) and a 3D HDR-to-2D LDR Mapping module (Sec. 3.4), and finally present our training strategies in Sec. 3.5.

3.1 Problem Setup

Inputs. Consider V *uncalibrated* views of a single 3D scene, given as context images $\{I_v\}_{v=1}^{V}$, $I_v \in \mathbb{R}^{H\times W\times 3}$, each captured at a known exposure time Δt_v. For convenience, we work in log-exposure space and define $\ell_v = \log_2 \Delta t_v$.

Outputs. InstantHDR jointly reconstructs the scene geometry and HDR appearance by predicting:
(a) HDR 3D Gaussians. A collection of G anisotropic 3D Gaussians

$$\left\{(\boldsymbol{\mu}_g,\, \sigma_g,\, \boldsymbol{r}_g,\, \boldsymbol{s}_g,\, \boldsymbol{c}_g^h)\right\}_{g=1}^{G}, \tag{1}$$

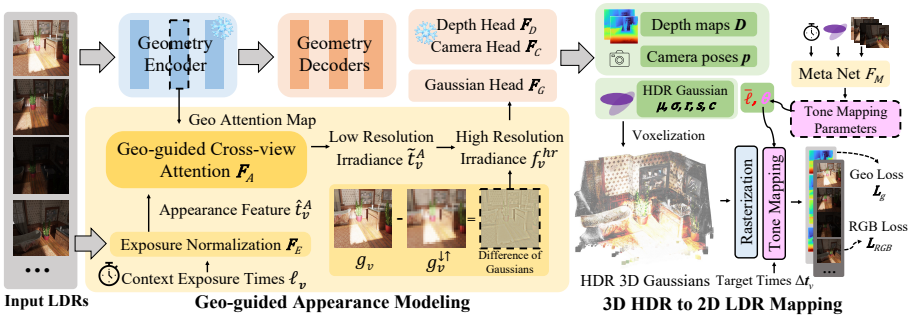


Fig. 2: Overview of InstantHDR. Given multi-exposure LDR images, the frozen geometry branch estimates depth and camera poses, while the appearance branch normalizes exposures (F_E), fuses cross-view irradiance via geometry-guided attention (F_A), and recovers pixel-level details via DoG upsampling. The Gaussian head F_G combines both branches to produce HDR 3D Gaussians. The Meta Net F_M predicts tone-mapping parameters for rendering LDR images at controllable exposures.

where each Gaussian is parameterized by a center position $\boldsymbol{\mu} \in \mathbb{R}^3$, an opacity $\sigma \in \mathbb{R}^+$, an orientation quaternion $\boldsymbol{r} \in \mathbb{R}^4$, an anisotropic scale $\boldsymbol{s} \in \mathbb{R}^3$, and an HDR color embedding $\boldsymbol{c}^h \in \mathbb{R}^{3\times(k+1)^2}$ parameterized as degree-k spherical-harmonic (SH) coefficients that encode *log-radiance*, following [2, 35].
(b) Camera parameters. Per-view parameters $\{p_v \in \mathbb{R}^9\}_{v=1}^V$, where p_v comprises a focal length, a 3-DoF rotation (axis-angle), a 3-DoF translation, and a 2-DoF principal-point offset.
(c) Scene-level attributes. A mid-exposure anchor $\bar{\ell} = \frac{1}{2}(\max_v \ell_v + \min_v \ell_v)$, serving as the reference exposure level, and the parameters $\boldsymbol{\theta}$ of a lightweight tonemapper that approximates the scene-specific camera response function (CRF).

Overall mapping. Formally, our model implements:

$$f_{\boldsymbol{\Theta}}\colon \{I_v, \ell_v\}_{v=1}^V \longmapsto \left\{(\boldsymbol{\mu}_g, \sigma_g, \boldsymbol{r}_g, \boldsymbol{s}_g, \boldsymbol{c}_g^h)\right\}_{g=1}^G \cup \{p_v\}_{v=1}^V \cup \{\bar{\ell}, \boldsymbol{\theta}\}. \quad (2)$$

3.2 Pipeline Overview

As illustrated in Fig. 2, our pipeline consists of two branches: a *geometry branch* that estimates scene structure from multi-view images, and an *appearance branch* that reconstructs HDR irradiance from exposure-inconsistent inputs. Given V uncalibrated multi-exposure LDR images, the geometry branch encodes them into high-dimensional features via a pretrained transformer and decodes depth maps and camera poses. The appearance branch—our core contribution—uses a *Geo-guided Appearance Modeling* module that leverages geometric correspondences from the geometry branch to fuse multi-exposure information into a coherent HDR representation. The outputs of both branches are combined by a Gaussian head to produce HDR 3D Gaussians, which are then converted to LDR via a learned tonemapper (Sec. 3.4).

Geometry Branch. The geometry branch provides the structural foundation for our pipeline. Following VGGT [63] and AnySplat [18], we adopt a pretrained alternating-attention transformer as the geometry backbone. Each image I_v is patchified into $N = \frac{HW}{p^2}$ tokens of dimension d using DINOv2 [50], where p=14 and d=1024. To each token sequence $\boldsymbol{t}_v^G \in \mathbb{R}^{N\times d}$, we prepend a learnable camera token $\boldsymbol{t}_v^{\text{cam}} \in \mathbb{R}^{1\times d}$ and four register tokens $\boldsymbol{t}_v^R \in \mathbb{R}^{4\times d}$. The combined tokens from all V views are processed by an L-layer alternating-attention transformer, where each layer applies frame-wise self-attention followed by global cross-view attention. Dedicated decoder heads for camera poses p_v and depth maps D_v are *frozen* together with the geometry encoder throughout training.

Gaussian Head. The Gaussian head combines information from both branches to predict HDR-aware Gaussian attributes. It takes the geometry tokens $\boldsymbol{t}_v^G$ and the high-resolution irradiance features $\boldsymbol{f}_v^{\text{hr}}$ produced by the Geo-guided Appearance Modeling module (Sec. 3.3) as input. The Gaussian head remains *trainable*, enabling it to output HDR-aware Gaussian attributes $\{\sigma_g, \boldsymbol{r}_g, \boldsymbol{s}_g, \boldsymbol{c}_g^h\}$.

3.3 Geo-guided Appearance Modeling

The frozen geometry branch provides reliable structure but cannot handle exposure-induced appearance inconsistency. We introduce the Geo-guided Appearance Modeling module, which mitigates this problem through three stages: (1) *Exposure Normalization* aligns inputs to a common reference level, (2) *Geo-guided Cross-view Attention* fuses irradiance features using geometric correspondences from the frozen backbone, and (3) *High-Resolution Upsampling* recovers pixel-level textures via high-frequency cues. The output features are decoded into log-radiance SH colors $\boldsymbol{c}_g^h$ for each Gaussian.

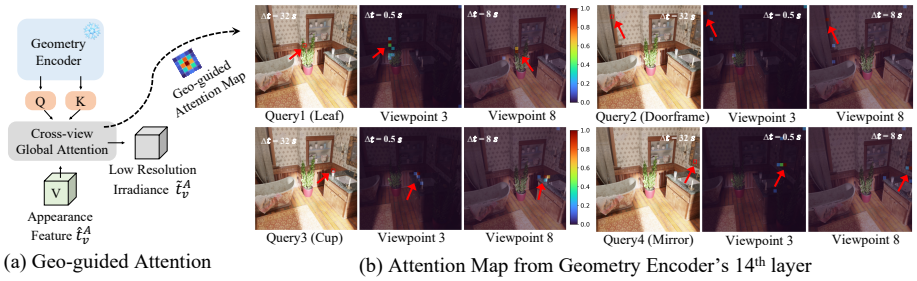
Exposure Normalization F_E**.** To fuse multi-exposure views, we first remove the exposure-induced brightness variation by normalizing all appearance features to a shared reference level. We define the relative log-exposure of each view as $\tilde{\ell}_v = \ell_v - \bar{\ell}$, where $\bar{\ell}$ is the mid-exposure anchor from Eq. (2), and encode $\tilde{\ell}_v$ into a d-dimensional embedding $\mathbf{e}_v$ via sinusoidal positional encoding [47]. Meanwhile, we extract per-view appearance tokens $\boldsymbol{t}_v^A \in \mathbb{R}^{N\times d}$ from each LDR image using a separate patch encoder with the same patch size p and dimension d as the geometry backbone. A FiLM layer [51] then predicts per-view affine parameters:

$$(\gamma_v,\, \beta_v) = \text{FiLM}\big(\mathbf{e}_v,\, \bar{\mathbf{a}}_v,\, \bar{\mathbf{a}}\big), \tag{3}$$

where $\bar{\mathbf{a}}_v = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{t}_{v,n}^A$ is the per-view feature mean and $\bar{\mathbf{a}} = \frac{1}{V}\sum_{v=1}^{V} \bar{\mathbf{a}}_v$ is the global scene summary. The appearance tokens are then modulated as:

$$\hat{\boldsymbol{t}}_v^A = \boldsymbol{t}_v^A \odot (1+\gamma_v) + \beta_v, \tag{4}$$

aligning all views to a shared irradiance level before cross-view fusion. Note that unlike methods that first linearize inputs via inverse gamma correction [85],



(a) Geo-guided Attention (b) Attention Map from Geometry Encoder's 14th layer

Fig. 3: Geo-guided Cross-view Attention. (a) The module reuses Q, K from the 14th frozen geometry encoder layer to guide appearance fusion. (b) Attention maps visualization shows that the it naturally and accurately matches query patches (red box) across views under large viewpoint and extreme exposure variations (Δt: 0.5–32s).

our model operates directly on camera-output LDR images (which are typically gamma-encoded), since our training objective reconstructs multi-exposure LDR images rather than linear HDR radiance (Sec. 3.5).

Geo-guided Cross-view Attention F_A**.** Different exposures capture complementary information: bright exposures reveal shadows while dark ones preserve highlights. Fusing them requires cross-view correspondences, which are challenging under large viewpoint and exposure changes. Interestingly, we observe that *the global attention maps in the frozen geometry encoder already encode reliable cross-view geometric correspondences, greatly benefiting our appearance modeling.* Therefore, we reuse these attention maps to guide appearance fusion. As shown in Fig. 3 (b), diverse elements such as leaves, cups, doorframes, and mirrors are accurately matched across views despite extreme exposure variations.

$$\tilde{\boldsymbol{t}}_v^A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)\hat{\boldsymbol{t}}_v^A. \tag{5}$$

High-Resolution Upsampling F_U**.** The irradiance features $\tilde{\boldsymbol{t}}_v^A$ operate at patch resolution $(\frac{H}{p} \times \frac{W}{p})$, losing high-frequency details critical for realistic appearance. To recover pixel-level textures while preserving the fused low-frequency irradiance, we adopt a Difference-of-Gaussians (DoG) guided upsampling strategy. We first encode each full-resolution LDR image into a feature map $\mathbf{g}_v \in \mathbb{R}^{d' \times H \times W}$ via a shallow CNN. A low-passed version $\mathbf{g}_v^{\downarrow\uparrow}$ is obtained by downsampling and bilinearly upsampling $\mathbf{g}_v$, yielding the high-frequency residual $\mathbf{g}_v - \mathbf{g}_v^{\downarrow\uparrow}$. This residual is then added to the bilinearly upsampled irradiance features to produce pixel-level irradiance features:

$$\boldsymbol{f}_v^{\mathrm{hr}} = \mathrm{Conv}\big(\mathrm{Up}(\tilde{\boldsymbol{t}}_v^A) + \mathrm{Conv}(\mathbf{g}_v - \mathbf{g}_v^{\downarrow\uparrow})\big) \in \mathbb{R}^{d \times H \times W}, \tag{6}$$

combining multi-view irradiance consensus with per-image structural detail.

HDR 3D Gaussian Prediction. We now merge the geometry and appearance branches to produce HDR 3D Gaussians. A DPT decoder [53] upsamples the geometry tokens $\boldsymbol{t}_v^G$ to pixel resolution, and the result is added to the irradiance features $\boldsymbol{f}_v^{\text{hr}}$. A lightweight CNN then regresses per-Gaussian opacity, orientation, scale, and log-radiance SH color:

$$\{\sigma_g,\, \boldsymbol{r}_g,\, \boldsymbol{s}_g,\, \boldsymbol{c}_g^h\} = F_G\big(\text{DPT}(\boldsymbol{t}_v^G) + \boldsymbol{f}_v^{\text{hr}}\big). \tag{7}$$

The Gaussian centers $\boldsymbol{\mu}_g$ are obtained by back-projecting the predicted depth maps D_v through the estimated camera poses p_v. The Gaussians are then voxelized [18] to reduce primitive count for efficient splatting.

3.4 3D HDR-to-2D LDR Mapping

Given HDR 3D Gaussians, rendering a photorealistic LDR view requires recovering the Camera Response Function (CRF) that maps scene irradiance to observed pixel values. Unlike optimization-based methods that overfit a per-scene MLP tonemapper [2, 35], our method seeks a *generalizable* tonemapper that adapts to different cameras without per-scene optimization. We achieve this goal via a **Meta Net** F_M that predicts the parameters of a lightweight tonemapper from scene context.

Tone Mapping Formulation. We convert HDR radiance to LDR in two steps. First, Gaussian splatting rasterizes the linear radiance into a per-pixel HDR image at view v:

$$\mathbf{H}_v = \mathcal{R}\big(\exp(\boldsymbol{c}_g^h),\, \sigma_g,\, \boldsymbol{r}_g,\, \boldsymbol{s}_g,\, \boldsymbol{\mu}_g;\, p_v\big) \;\in\; \mathbb{R}^{H\times W\times 3}, \tag{8}$$

where $\mathcal{R}(\cdot)$ denotes the differentiable rasterization with alpha-weighted blending [26] in linear radiance space, p_v is the camera pose, and the exp converts log-radiance SH colors $\boldsymbol{c}_g^h$ back to linear radiance before blending.

Second, following the log-domain CRF model [7], a learned tonemapper $g_{\boldsymbol{\theta}}$ maps the log-irradiance to $[0, 1]$ LDR values:

$$\mathbf{L}_v(\ell) = g_{\boldsymbol{\theta}}\big(\log \mathbf{H}_v + (\ell - \bar{\ell}) \cdot \log 2\big), \tag{9}$$

where $\mathbf{L}_v(\ell) \in \mathbb{R}^{H\times W\times 3}$ is the rendered LDR image at view v under target log-exposure ℓ, $\bar{\ell}$ is the mid-exposure anchor from Eq. (2), and the term $(\ell - \bar{\ell}) \cdot \log 2$ adjusts the irradiance to the desired exposure level. Here $g_{\boldsymbol{\theta}}$ is a two-layer MLP with hidden dimension h (input: $3 \to$ hidden: h with ReLU $\to$ output: 3 with sigmoid), acting as a learnable inverse CRF. Rather than learning a fixed $\boldsymbol{\theta}$, we predict *scene-specific* parameters via the Meta Net, enabling adaptation to different cameras and tone curves without per-scene optimization.

Meta Net F_M**.** The Meta Net infers scene-specific tonemapper parameters $\boldsymbol{\theta}$ in a single forward pass, enabling $g_{\boldsymbol{\theta}}$ to reproduce the original camera's tone curve

without per-scene optimization. For brevity, we denote the full set of pixel-level Gaussians (before voxelization) as:

$$\mathcal{G} = \left\{(\boldsymbol{\mu}_g,\, \sigma_g,\, \boldsymbol{r}_g,\, \boldsymbol{s}_g,\, \boldsymbol{c}_g^h)\right\}_{g=1}^{G}, \quad G = V \times H \times W. \tag{10}$$

The Meta Net takes three inputs: (i) the full-resolution LDR features $\mathbf{g}_v$ from the upsampling CNN (Eq. (6)), (ii) the per-view exposure embeddings $\{\mathbf{e}_v\}_{v=1}^{V}$, and (iii) the predicted HDR Gaussians $\mathcal{G}$. These are concatenated and compressed by a strided convolutional encoder and then globally pooled across all spatial and views dimensions to produce a scene-level descriptor $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = F_M\big(\{\mathbf{g}_v\},\, \{\mathbf{e}_v\},\, \mathcal{G}\big) \;\in\; \mathbb{R}^{d_\theta}, \tag{11}$$

where d_θ encodes all weights and biases of the two-layer tonemapper $g_{\boldsymbol{\theta}}$.

3.5 Training Strategies

Training Objective. InstantHDR is trained end-to-end without any 3D or HDR supervision, using only multi-view LDR images with known exposure times. The geometry encoder and its decoder heads remain frozen; only the appearance branch, the Gaussian head, and the Meta Net are optimized.

During training, the target views are same with the context views, i.e., the model predicts HDR Gaussians $\mathcal{G}$ and tonemapper parameters $\boldsymbol{\theta}$ from $\{I_v, \ell_v\}_{v=1}^{V}$ and is supervised by rendering back to the same views and exposures via Eqs. (8)–(9). At test time, the target view and exposure can differ from the context set, enabling novel-view synthesis at arbitrary exposures. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{RGB}} + \lambda_g\, \mathcal{L}_g. \tag{12}$$

The photometric loss compares rendered and ground-truth LDR images:

$$\mathcal{L}_{\text{RGB}} = \frac{1}{V}\sum_{v=1}^{V}\Big[\text{MSE}\big(I_v,\, \mathbf{L}_v(\ell_v)\big) + \lambda_{\text{perc}} \cdot \mathcal{L}_{\text{perc}}\big(I_v,\, \mathbf{L}_v(\ell_v)\big)\Big]. \tag{13}$$

The geometry consistency loss enforces alignment between the depth maps D_v from the frozen DPT head and the rendered depth maps $\hat{D}_v$ from the predicted Gaussians. Since D_v can be unreliable in challenging regions (*e.g.*, sky or reflective surfaces), we utilize the jointly learned confidence map C_v^D and apply supervision only to the top-N% most confident pixels:

$$\mathcal{L}_g = \frac{1}{V}\sum_{v=1}^{V}\big(D_v[M_v] - \hat{D}_v[M_v]\big)^2, \tag{14}$$

where M_v is a binary mask corresponding to the top-N quantile of C_v^D; we set N=30 in all experiments. Our model learns HDR representations implicitly by correctly reconstructing LDR images across different exposure times.

Table 1: Comparison of existing HDR datasets for novel view synthesis. "Pano." denotes panoramic 360° images. "Multi-Exp." indicates multi-exposure LDR images, with the number of exposures in parentheses.

Dataset	Source	#Scenes	Type	HDR GT	Multi-Exp.	Depth/Normal
HDR-NeRF [15]	Blender + Real	8 + 4	Object/Indoor	✓	✓ (5)	×
HDR-Plenoxels [21]	Blender + Real	5 + 4	Indoor	×	✓ (3)	×
Pano-NeRF [42]	Blender + Replica + Real	5 + 8 + 3	Indoor (Pano.)	✓	✓ (9)	✓
PanDORA [5]	Real capture (360°)	14	Indoor (Pano.)	✓	✓ (2)	×
HDR-Pretrain (Ours)	Blender	**168**	Indoor	✓	✓ (5)	✓

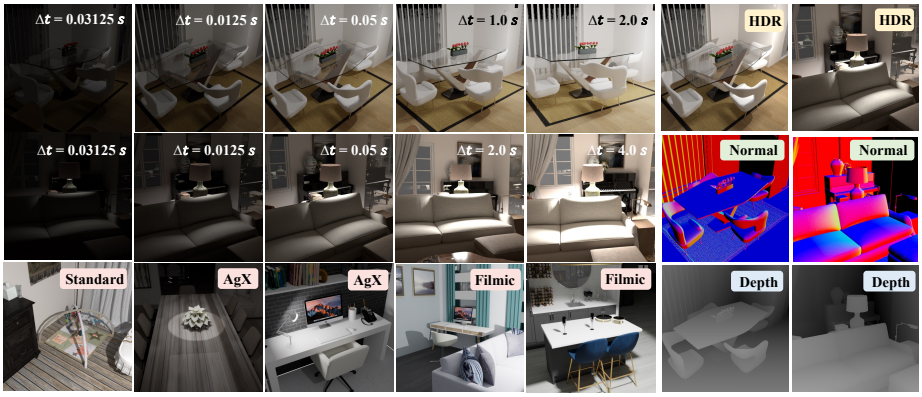


Fig. 4: Examples from our HDR-Pretrain dataset. Each scene includes multi-view, multi-exposure LDR images at varying Δt, 32-bit HDR ground truth, depth and normal maps, rendered under diverse tone-mapping operators (Standard, AgX, Filmic).

Post-Optimization. We also can refine the predicted Gaussians and camera parameters via post optimization. After pruning low-opacity Gaussians ($\sigma < 0.01$), we minimize a combination of MSE and SSIM losses between rendered and input images for 1K iterations, back-propagating gradients through all Gaussian attributes, and tonemapper parameters. The learning rates are: 1.6e-4 (position), 5e-3 (scale), 1e-3 (rotation), 5e-2 (opacity), and 2.5e-3 (color).

4 Experiments

Pretraining Dataset. HDR datasets with multi-view LDR images remain extremely scarce. As shown in Tab. 1, existing benchmarks contain only a handful of scenes, far from sufficient for large-scale pretraining. To bridge this gap, we construct **HDR-Pretrain**, in Fig. 4, a large-scale synthetic dataset of 168 photorealistic indoor scenes rendered in Blender. The 3D assets are sourced from HSSD [28], an open-source collection of realistic interiors originally built for embodied AI research. Following HDR-NeRF [15], we sample viewpoints on a 5×7 grid with 2.5° / 5° angular steps per scene and render 32-bit HDR images via Cycles path tracing at 448×448 resolution. Each view is paired with 5 exposure-bracketed LDR images under a randomly chosen tone-mapping operator, as well

as depth and normal maps. We randomly apply one of three tone-mapping operators (AgX, Filmic, Standard) per scene to increase data diversity.

Evaluation Dataset. We mainly evaluate on the HDR-NeRF benchmark [15], which contains 8 synthetic and 4 real indoor scenes, each captured from 35 viewpoints at 5 exposure levels $\{t_1, t_2, t_3, t_4, t_5\}$. Following the standard protocol [15], 18 views with one exposure drawn from $\{t_1, t_3, t_5\}$ are used as input, while the remaining views are held out for evaluation. For clarity, we report the average LDR metrics across all 5 exposures rather than separating observed (LDR-OE) and novel (LDR-NE) exposures. We further test under sparse-view settings with only 4 or 8 input views.

Implementation Details. We build upon AnySplat [18] with its backbone frozen, using a voxel size of ϵ=0.002. Training uses AdamW with cosine scheduling, peak learning rate 2×10^{-4}, 1K warmup, and runs for 30K iterations in bf16 precision on 8 NVIDIA A6000 GPUs for ∼2 days. Input resolution is 448×448 with random cropping and flipping augmentation. We set λ_{perc}=0.05 and λ_{g}=0.1, sampling 2∼10 context views per iteration. Due to the domain gap between real and our synthetic HDR-Pretrain scenes, we finetune on HDR-Plenoxels 4 real scenes before evaluating on HDR-NeRF real scenes, ensuring the testing scenes is always unseen. Following pose-free methods [76], we perform test-time pose alignment for evaluation and post optimization.

Evaluation Metrics. All images are resized to 448×448 for fair comparison. We report PSNR, SSIM, LPIPS [83], and reconstruction time for quantitative and efficiency comparison. Following HDR-NeRF [15], HDR results are evaluated in the tone-mapped domain using the μ-law [23], with 99th-percentile normalization to suppress extreme HDR values.

4.1 Quantitative Results

LDR Comparisons on Zero-shot Inference. We first evaluate zero-shot results on HDR-NeRF [15] scenes. AnySplat [18] ignores exposure inconsistency, leading to severely degraded results. As shown in Tab. 2, InstantHDR consistently outperforms AnySplat by large margins (*e.g.*, +5.65 dB on real scenes, +8.07 dB on synthetic scenes with 8 views) with negligible increases of reconstruction time, showing strong generalization to exposure-inconsistent inputs.

LDR Comparisons on Optimization. We further compare against state-of-the-art optimization-based methods, HDR-GS [2] and GaussianHDR [35]. In this setting, we post-optimize the Gaussians and tone-mapping parameters produced by InstantHDR and AnySplat for 1K iterations, denoted as InstantHDR_1K and AnySplat_1K. Under the challenging 4-view sparse setting on real scenes, InstantHDR_1K achieves 22.16 dB PSNR and 0.762 SSIM, surpassing GaussianHDR by +2.90 dB and +0.071 in SSIM, demonstrating that the diverse geometric priors from feed-forward foundation models greatly benefit sparse-view reconstruction. Under the denser 18-view setting, our method remains competitive on real scenes while showing a modest gap on synthetic scenes. Notably, InstantHDR_1K requires only ∼30–40 seconds per scene, roughly **20**× faster

Table 2: Quantitative LDR comparison on HDR-NeRF [15] real and synthetic scenes with varying numbers of input views. We compare our method, as well as its variants with 1K iters of post-optimization (Ours_1K), against AnySplat [18] HDR-GS [2] and GaussianHDR [35]. Time denotes per-scene reconstruction time in seconds.

Mode	Method	4 Views				8 Views				18 Views			
		PSNR↑	SSIM↑	LPIPS↓	Time(s)↓	PSNR↑	SSIM↑	LPIPS↓	Time(s)↓	PSNR↑	SSIM↑	LPIPS↓	Time(s)↓
HDR-NeRF Real Dataset [15]													
Zero-shot	AnySplat [18]	12.10	0.517	0.497	**0.973**	13.30	0.569	0.436	**1.180**	13.91	0.600	0.403	**2.139**
	InstantHDR (Ours)	**18.44**	**0.721**	**0.269**	1.033	**18.95**	**0.724**	**0.269**	1.582	**19.48**	**0.745**	**0.257**	2.512
Optimization	HDR-GS [2]	15.40	0.622	0.334	872	23.02	0.791	0.121	736	27.42	0.893	0.047	815
	GaussianHDR [35]	19.26	0.691	0.270	1833	24.96	**0.854**	**0.068**	1816	**29.36**	0.929	**0.024**	1891
	AnySplat_1K [18]	11.84	0.486	0.468	**30**	13.63	0.580	0.372	**33**	14.86	0.689	0.266	40
	InstantHDR_1K (Ours)	**22.16**	**0.762**	**0.259**	32	**25.32**	0.852	0.160	40	29.19	**0.931**	0.086	**39**
HDR-NeRF Syn Dataset [15]													
Zero-shot	AnySplat [18]	13.98	0.525	0.400	-	14.51	0.573	0.378	-	15.27	0.534	0.327	-
	InstantHDR (Ours)	**21.76**	**0.728**	**0.172**		**22.58**	**0.785**	**0.138**		**22.59**	**0.830**	**0.115**	
Optimization	HDR-GS [2]	24.26	0.711	0.210	-	30.60	0.867	0.086	-	29.93	0.917	0.061	-
	GaussianHDR [35]	21.62	0.646	0.224		**34.49**	**0.924**	**0.026**		**38.63**	**0.969**	**0.009**	
	AnySplat_1K [18]	14.01	0.526	0.347		15.42	0.656	0.248		16.39	0.624	0.221	
	InstantHDR_1K (Ours)	**27.63**	**0.825**	**0.137**		32.75	0.922	0.061		35.99	0.965	0.037	

Table 3: (a) Quantitative HDR comparison on HDR-NeRF [15] synthetic scenes with 8 input views, evaluated in the μ-law tone-mapped domain. (b) Ablation study of InstantHDR. Zero-shot results on HDR-NeRF [15] real scenes with 8 input views.

(a) HDR Comparision on HDR-NeRF [15] syn scenes.

Mode	Method	PSNR↑	SSIM↑	LPIPS↓	Time(s)↓
Zero-shot	AnySplat	8.93	0.595	0.416	**1.180**
	InstantHDR (Ours)	**15.29**	**0.772**	**0.140**	1.582
Optimization	HDR-GS	27.69	0.871	0.090	736
	GaussianHDR	**31.62**	0.887	**0.037**	1816
	AnySplat_1K	9.52	0.678	0.268	**33**
	InstantHDR_1K (Ours)	27.55	**0.899**	0.076	40

(b) Ablation on HDR-NeRF [15] real scenes.

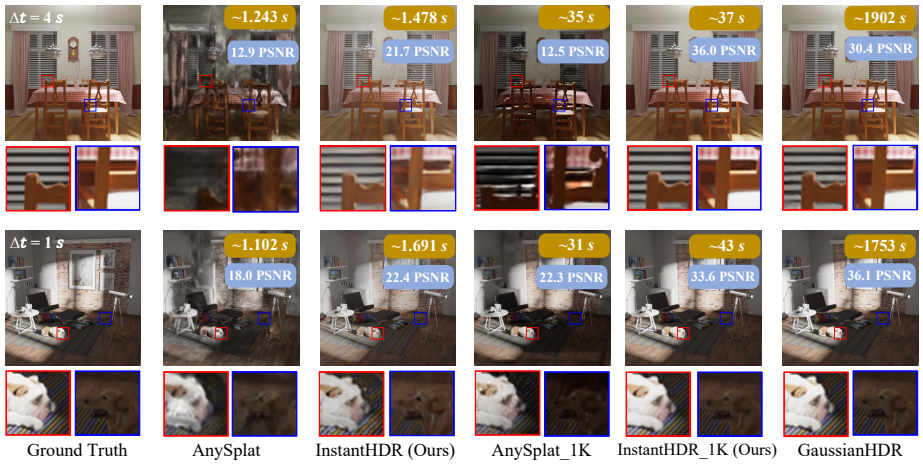
Method	PSNR↑	SSIM↑	LPIPS↓
w/o Meta Net	16.32	0.699	0.289
w/o Exposure Norm	13.72	0.693	0.278
w/o Cross-view Attn	17.63	0.702	0.277
w/o Upsampling	**19.20**	0.718	0.386
Ours	18.95	**0.724**	**0.269**

than HDR-GS and **50**× faster than GaussianHDR, attributing to the good feed-forward initialization and the ability to skip the costly iterative densification.

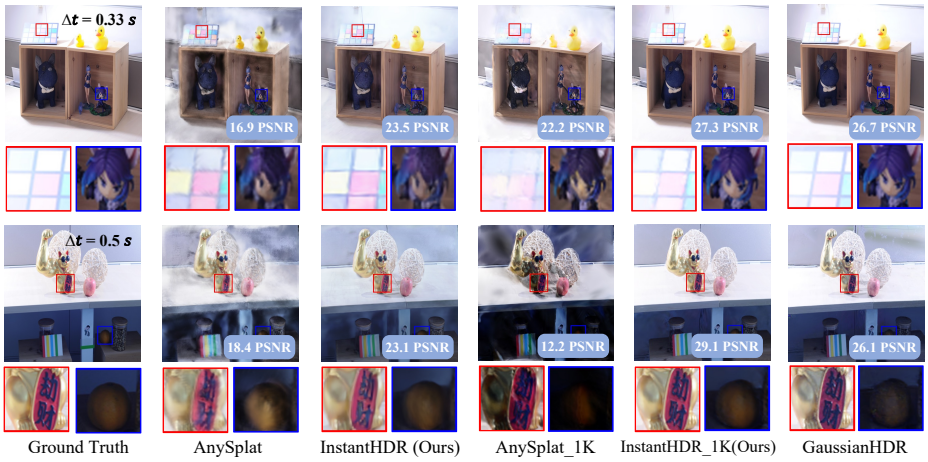
HDR Comparisons. Tab. 3 (a) reports HDR results on HDR-NeRF synthetic scenes. Although never supervised with HDR ground truth, InstantHDR implicitly recovers HDR from multi-exposure LDR inputs, outperforming AnySplat by +6.36 dB in PSNR in zero-shot mode. With 1K steps of post-optimization, InstantHDR_1K reaches 27.55 dB PSNR and 0.899 SSIM, achieving comparable performance to HDR-GS (27.69 dB) while surpassing GaussianHDR in SSIM. The remaining gap with GaussianHDR in PSNR is likely due to its dedicated 3D–2D dual-branch tone-mapping design, whereas both our method and HDR-GS adopt a simpler single tone mapping branch. We believe that incorporating more advanced tone-mapping modules is a promising avenue for future works.

4.2 Qualitative Results

LDR Novel View Rendering. In Fig. 5, AnySplat struggles in exposure variations, while GaussianHDR is time-consuming. Our InstantHDR generalizes well to both synthetic and real-world scenes, renders clean, exposure-controllable LDR views in seconds, and achieves competitive quality after 1K post-optimization.



(a) LDR visual comparisions on HDR-NeRF [15] synthetic scenes.



(b) LDR visual comparisions on HDR-NeRF [15] real scenes.

Fig. 5: LDR visual comparisons. Feed-forward methods [18] fail on multi-exposure inputs, while optimization-based methods [35] require ∼2K seconds per scene. Our InstantHDR achieves competitive quality in under 40s. Yellow/blue tags denote reconstruction time/PSNR.

HDR Novel View Rendering. As shown in Fig. 6, zero-shot HDR outputs from feed-forward models (AnySplat & InstantHDR) appear overly bright, as extreme radiance values are hard to predict accurately in a single-foward, so inflate the average brightness after normalization. We see this as an open challenge for feed-forward HDR models. After 1K post-optimization, this issue is largely alleviated, with our InstantHDR producing results similar to GaussianHDR.

Fig. 6: HDR visual comparisons. After 1K post-optimization, our InstantHDR produces HDR results comparable to time-consuming GaussianHDR.

Fig. 7: Ablation visualization. Attn.=cross-view attention, Exp.=exposure normalization, Up.=upsampling. Removing Attn. causes wall ghosting; removing Exp. shifts overall brightness; removing Up. produces blurry results.

4.3 Ablation Study

As shown in Tab. 3 (b) and Fig. 7, each component plays a critical role. Removing the MetaNet makes training unstable, as the model cannot adapt to varying camera response functions. Eliminating exposure normalization leads to the largest performance degradation, since inconsistent brightness across views disrupts feature fusion. Disabling cross-view attention introduces ghosting artifacts on smooth surfaces such as walls. Finally, removing the upsampling module preserves coarse structure but loses fine details, resulting in blurry outputs.

5 Conclusion

We present InstantHDR, the first feed-forward framework for HDR novel view synthesis from uncalibrated multi-exposure LDR images. By leveraging geometry-guided cross-view attention for exposure-robust appearance fusion and a meta-network for scene-adaptive tone mapping, InstantHDR reconstructs HDR scenes in few seconds. We also introduce HDR-Pretrain, a 168-scene synthetic dataset to address the data scarcity for feed-forward HDR pretraining. Experiments demonstrate that InstantHDR achieves competitive quality to optimization-based methods while being orders of magnitude faster. We hope this InstantHDR can inspire more ideas on real-time 3D HDR reconstruction.

References

1. Cai, Y., Liang, Y., Wang, J., Wang, A., Zhang, Y., Yang, X., Zhou, Z., Yuille, A.: Radiative gaussian splatting for efficient x-ray novel view synthesis. In: ECCV (2024) 4
2. Cai, Y., Xiao, Z., Liang, Y., Qin, M., Zhang, Y., Yang, X., Liu, Y., Yuille, A.L.: Hdr-gs: Efficient high dynamic range novel view synthesis at 1000x speed via gaussian splatting. NeurIPS **37**, 68453–68471 (2024) 2, 4, 5, 8, 11, 12
3. Chen, R., Zheng, B., Zhang, H., Chen, Q., Yan, C., Slabaugh, G., Yuan, S.: Improving dynamic hdr imaging with fusion transformer. In: AAAI (2023) 3
4. Chen, Z., Yang, J., Yang, H.: Pref3r: Pose-free feed-forward 3d gaussian splatting from variable-length image sequence. arXiv preprint arXiv:2411.16877 (2024) 4
5. Dastjerdi, M.R.K., Tanguay-Gaudreau, D., Fortier-Chouinard, F., Hold-Geoffroy, Y., Demers, C., Kalantari, N., Lalonde, J.F.: Pandora: Casual hdr radiance acquisition for indoor scenes. arXiv preprint arXiv:2407.06150 (2024) 10
6. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: SIGGRAPH (1997) 3
7. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 643–652 (2023) 8
8. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: Hdr image reconstruction from a single exposure using deep cnns. ACM TOG (2017) 3
9. Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., Zhang, B., Dai, B.: Generative diffusion prior for unified image restoration and enhancement. In: CVPR (2023) 4
10. Grosch, T., et al.: Fast and robust high dynamic range image generation with camera and object movement. Vision, Modeling and Visualization, RWTH Aachen (2006) 3
11. Hong, S., Jung, J., Shin, H., Han, J., Yang, J., Luo, C., Kim, S.: Pf3plat: Pose-free feed-forward 3d gaussian splatting. arXiv preprint arXiv:2410.22128 (2024) 4
12. Hu, S., Liu, Z.: Gauhuman: Articulated gaussian splatting from monocular human videos. arXiv preprint arXiv: (2023) 4
13. Hu, T., Sarkar, K., Liu, L., Zwicker, M., Theobalt, C.: Egorenderer: Rendering human avatars from egocentric camera images. In: ICCV. pp. 14528–14538 (2021) 2
14. Huang, S., Gojcic, Z., Wang, Z., Williams, F., Kasten, Y., Fidler, S., Schindler, K., Litany, O.: Neural lidar fields for novel view synthesis. In: ICCV. pp. 18236–18246 (2023) 2
15. Huang, X., Zhang, Q., Feng, Y., Li, H., Wang, X., Wang, Q.: Hdr-nerf: High dynamic range neural radiance fields. In: CVPR. pp. 18398–18408 (2022) 2, 3, 4, 10, 11, 12, 13
16. Jacobs, K., Loscos, C., Ward, G.: Automatic high-dynamic range image generation for dynamic scenes. IEEE Computer Graphics and Applications (2008) 3
17. Jiang, H., Jiang, Z., Zhao, Y., Huang, Q.: Leap: Liberate sparse-view 3d modeling from camera poses. arXiv preprint arXiv:2310.01410 (2023) 4
18. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM TOG **44**(6), 1–16 (2025) 2, 3, 6, 8, 11, 12, 13
19. Jiang, Y., Tu, J., Liu, Y., Gao, X., Long, X., Wang, W., Ma, Y.: Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. arXiv preprint arXiv:2311.17977 (2023) 4

20. Jin, H., Li, Y., Luan, F., Xiangli, Y., Bi, S., Zhang, K., Xu, Z., Sun, J., Snavely, N.: Neural gaffer: Relighting any object via diffusion. In: NeurIPS (2024) 3
21. Jun-Seong, K., Yu-Ji, K., Ye-Bin, M., Oh, T.H.: Hdr-plenoxels: Self-calibrating high dynamic range radiance fields. In: ECCV. pp. 384–401. Springer (2022) 3, 4, 10
22. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. ACM ToG (2017) 3
23. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**(4), 144–1 (2017) 11
24. Keetha, N., Karhade, J., Jatavallabhula, K.M., Yang, G., Scherer, S., Ramanan, D., Luiten, J.: Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. arXiv preprint arXiv:2312.02126 (2023) 4
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (2023) 4
26. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG **42**(4), 139–1 (2023) 1, 8
27. Khan, Z., Khanna, M., Raman, S.: Fhdr: Hdr image reconstruction from a single ldr image using feedback network. In: IEEE Global Conference on Signal and Information Processing (2019) 3
28. Khanna, M., Mao, Y., Jiang, H., Haresh, S., Shacklett, B., Batra, D., Clegg, A., Undersander, E., Chang, A.X., Savva, M.: Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In: CVPR. pp. 16384–16393 (2024) 10
29. Kim, J., Lee, S., Kang, S.J.: End-to-end differentiable learning to hdr image synthesis for multi-exposure images. In: AAAI (2021) 3
30. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. NeurIPS **35**, 23311–23330 (2022) 2
31. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: Hugs: Human gaussian splats. arXiv preprint arXiv:2311.17910 (2023) 4
32. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV. pp. 71–91. Springer (2024) 4
33. Liang, Y., Yang, X., Lin, J., Li, H., Xu, X., Chen, Y.: Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. arXiv preprint arXiv:2311.11284 (2023) 4
34. Liang, Z., Zhang, Q., Feng, Y., Shan, Y., Jia, K.: Gs-ir: 3d gaussian splatting for inverse rendering. arXiv preprint arXiv:2311.16473 (2023) 4
35. Liu, J., Kong, L., Li, B., Xu, D.: Gausshdr: High dynamic range gaussian splatting via learning unified 3d and 2d local tone mapping. In: CVPR. pp. 5991–6000 (2025) 2, 4, 5, 8, 11, 12, 13
36. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM TOG **40**(6), 1–16 (2021) 2
37. Liu, S., Zhang, X., Zhang, Z., Zhang, R., Zhu, J.Y., Russell, B.: Editing conditional radiance fields. In: ICCV. pp. 5773–5783 (2021) 2
38. Liu, X., Zhan, X., Tang, J., Shan, Y., Zeng, G., Lin, D., Liu, X., Liu, Z.: Humangaussian: Text-driven 3d human generation with gaussian splatting. arXiv preprint arXiv:2311.17061 (2023) 4
39. Liu, Y., Dong, S., Wang, S., Yin, Y., Yang, Y., Fan, Q., Chen, B.: Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In: CVPR. pp. 16651–16662 (2025) 4

40. Liu, Z., Lin, W., Li, X., Rao, Q., Jiang, T., Han, M., Fan, H., Sun, J., Liu, S.: Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In: CVPRW (2021) 3
41. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: ECCV (2022) 3
42. Lu, Z., Zheng, Q., Shi, B., Jiang, X.: Pano-nerf: Synthesizing high dynamic range novel views with geometry from sparse low dynamic range panoramic images. In: AAAI. pp. 3927–3935 (2024) 10
43. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713 (2023) 4
44. Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian splatting slam. arXiv preprint arXiv:2312.06741 (2023) 4
45. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: Pacific Conference on Computer Graphics and Applications (2007) 3
46. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T.: Nerf in the dark: High dynamic range view synthesis from noisy raw images. In: CVPR (2022) 4
47. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 1, 6
48. Murai, R., Dexheimer, E., Davison, A.J.: Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In: CVPR. pp. 16695–16705 (2025) 4
49. Nazarczuk, M., Catley-Chandar, S., Leonardis, A., Pellitero, E.P.: Self-supervised hdr imaging from motion and exposure cues. arXiv preprint arXiv:2203.12311 (2022) 4
50. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 6
51. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018) 6
52. Prabhakar, K.R., Senthil, G., Agrawal, S., Babu, R.V., Gorthi, R.K.S.S.: Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting. In: CVPR (2021) 4
53. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021) 8
54. Smart, B., Zheng, C., Laina, I., Prisacariu, V.A.: Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. arXiv preprint arXiv:2408.13912 (2024) 4
55. Song, J.W., Park, Y.I., Kong, K., Kwak, J., Kang, S.J.: Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In: ECCV (2022) 3
56. Sun, J., Wang, X., Shi, Y., Wang, L., Wang, J., Liu, Y.: Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. ACM TOG **41**(6), 1–10 (2022) 2
57. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: CVPR. pp. 8248–8258 (2022) 2
58. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023) 4

59. Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., Yan, Z.: Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In: CVPR. pp. 5283–5293 (2025) 4
60. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: The state of the art in hdr deghosting: A survey and evaluation. In: Computer Graphics Forum (2015) 3
61. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: ICCV. pp. 9065–9076 (2023) 2
62. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. arXiv preprint arXiv:2408.16061 (2024) 4
63. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025) 2, 4, 6
64. Wang, P., Tan, H., Bi, S., Xu, Y., Luan, F., Sunkavalli, K., Wang, W., Xu, Z., Zhang, K.: Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. arXiv preprint arXiv:2311.12024 (2023) 4
65. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. arXiv preprint arXiv:2501.12387 (2025) 4
66. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (2024) 4
67. Ward, G., Reinhard, E., Debevec, P.: High dynamic range imaging & image-based lighting. In: SIGGRAPH (2008) 3
68. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023) 4
69. Xie, T., Zong, Z., Qiu, Y., Li, X., Feng, Y., Yang, Y., Jiang, C.: Physgaussian: Physics-integrated 3d gaussians for generative dynamics. arXiv preprint arXiv:2311.12198 (2023) 4
70. Yan, C., Qu, D., Wang, D., Xu, D., Wang, Z., Zhao, B., Li, X.: Gs-slam: Dense visual slam with 3d gaussian splatting. arXiv preprint arXiv:2311.11700 (2023) 4
71. Yan, Q., Zhang, S., Chen, W., Tang, H., Zhu, Y., Sun, J., Van Gool, L., Zhang, Y.: Smae: Few-shot learning for hdr deghosting with saturation-aware masked autoencoders. In: CVPR (2023) 4
72. Yan, Q., Zhu, Y., Zhang, Y.: Robust artifact-free high dynamic range imaging of dynamic scenes. Multimedia Tools and Applications (2019) 3
73. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. arXiv preprint arXiv:2501.13928 (2025) 4
74. Yang, Z., Yang, H., Pan, Z., Zhu, X., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023) 4
75. Yang, Z., Chai, Y., Anguelov, D., Zhou, Y., Sun, P., Erhan, D., Rafferty, S., Kretzschmar, H.: Surfelgan: Synthesizing realistic sensor data for autonomous driving. In: CVPR. pp. 11118–11127 (2020) 2
76. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. arXiv preprint arXiv:2410.24207 (2024) 4, 11
77. Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. arXiv preprint arXiv:2310.08529 (2023) 4
78. Yu, Y., Wang, H., Luo, T., Fan, H., Zhang, L.: Magic: Multi-modality guided image completion. In: ICLR (2024) 3

79. Yu, Y., Zeng, Z., Hua, H., Fu, J., Luo, J.: Promptfix: You prompt and we fix the photo. In: NeurIPS (2024) 3
80. Yuan, Y.J., Sun, Y.T., Lai, Y.K., Ma, Y., Jia, R., Gao, L.: Nerf-editing: geometry editing of neural radiance fields. In: CVPR. pp. 18353–18364 (2022) 2
81. Yugay, V., Li, Y., Gevers, T., Oswald, M.R.: Gaussian-slam: Photo-realistic dense slam with gaussian splatting. arXiv preprint arXiv:2312.10070 (2023) 4
82. Zha, R., Lin, T.J., Cai, Y., Cao, J., Zhang, Y., Li, H.: R^2-gaussian: Rectifying radiative gaussian splatting for tomographic reconstruction. In: NeurIPS (2024) 4
83. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018) 11
84. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. arXiv preprint arXiv:2502.12138 (2025) 4
85. Zhang, Z., Wang, H., Liu, S., Wang, X., Lei, L., Zuo, W.: Self-supervised high dynamic range imaging with multi-exposure images in dynamic scenes. In: ICLR (2024) 4, 6
86. Zheng, J., Jang, Y., Papaioannou, A., Kampouris, C., Potamias, R.A., Papantoniou, F.P., Galanakis, E., Leonardis, A., Zafeiriou, S.: Ilsh: The imperial light-stage head dataset for human head view synthesis. In: ICCV. pp. 1112–1120 (2023) 2
87. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling. In: CVPR. pp. 15893–15903 (2022) 2