Generalizable Sparse-View 3D Reconstruction from Unconstrained Images

Vinayak Gupta[1] Chih-Hao Lin[2] Shenlong Wang[2] Anand Bhattad[3] Jia-Bin Huang[1]

[1]University of Maryland, College Park [2]University of Illinois Urbana-Champaign [3]Johns Hopkins University

https://genwildsplat.github.io/

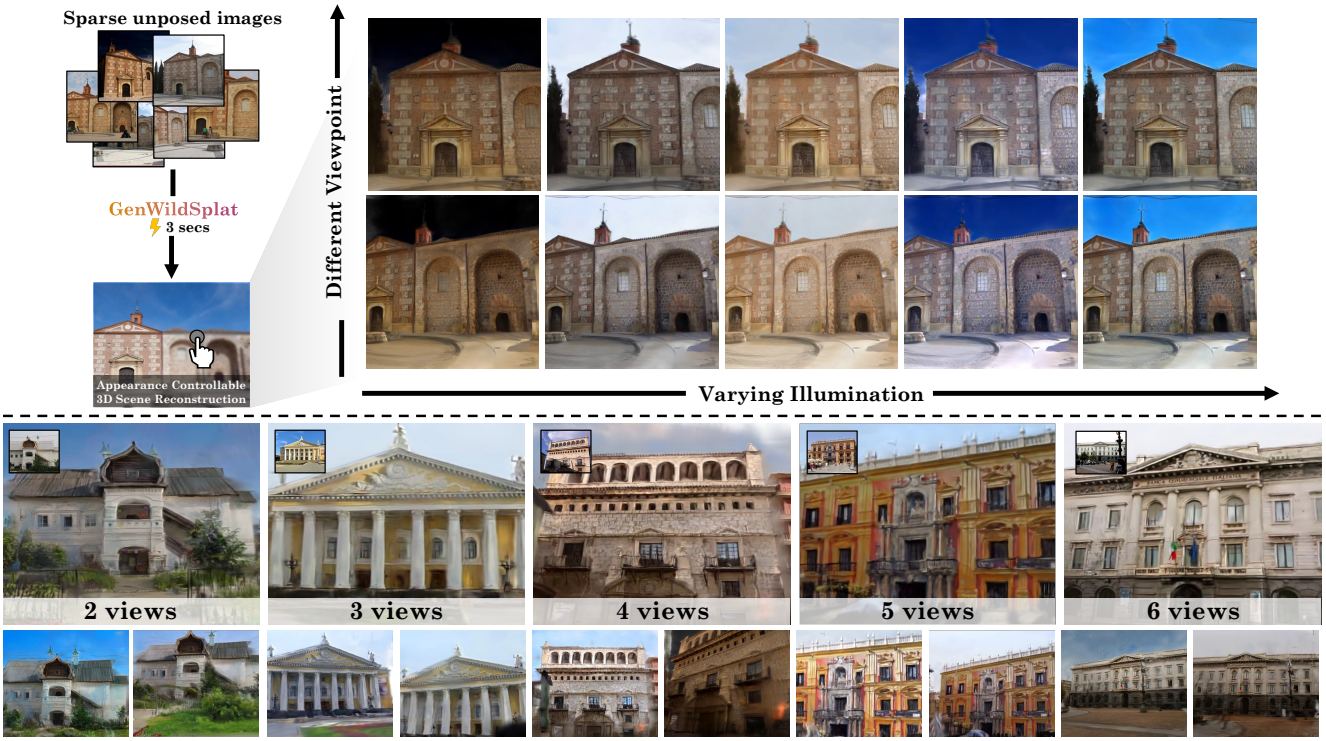

Figure 1. **GenWildSplat** reconstructs 3D scenes from sparse, unposed images with varying illumination and transient objects in a single 3-second feed-forward pass, and no per-scene optimization is required. Given 2–6 input views, our method predicts novel views under target lighting conditions while handling occlusions. **Top:** Novel-view synthesis under different lighting from the same sparse inputs, demonstrating appearance control. **Bottom:** Reconstruction quality across varying input sparsity (2–6 views), showing view-consistent rendering even with minimal observations. In each block, the top-left image (inset) is an input view; the remaining images are novel-view predictions under novel lighting. All scenes are unseen during training, demonstrating strong generalization to real-world environments.

Abstract

*Reconstructing 3D scenes from sparse, unposed images remains challenging under real-world conditions with varying illumination and transient occlusions. Existing methods rely on scene-specific optimization using appearance embeddings or dynamic masks, which requires extensive per-scene training and fails under sparse views. Moreover, evaluations on limited scenes raise questions about generalization. We present **GenWildSplat**, a feed-forward framework for sparse-view outdoor reconstruction that requires no per-scene optimization. Given unposed internet images, GenWildSplat predicts depth, camera parameters, and 3D Gaussians in a canonical space using learned geometric priors. An appearance adapter modulates appearance for target lighting conditions, while semantic segmentation handles transient objects. Through curriculum learning on synthetic and real data, GenWildSplat generalizes across diverse illumination and occlusion patterns. Evaluations on PhotoTourism and MegaScenes benchmark demonstrate state-of-the-art feed-forward rendering quality, achieving real-time inference without test-time optimization.*

1. Introduction

Reconstructing 3D scenes from 2D images is crucial for applications such as AR/VR and navigation [25, 56]. Extending these techniques to in-the-wild imagery remains challenging due to three factors: (1) Internet photos show wide lighting variations across time and seasons, (2) handheld captures contain transient occluders like tourists or vehicles that must be excluded, and (3) real-world scenes often provide sparse viewpoints, unlike curated multi-view datasets. Effective reconstruction requires disentangling static scene content from dynamic lighting and transient objects. Prior NeRF [4, 31, 34] and Gaussian Splatting [8, 14, 16, 43, 44, 48] methods rely on *per-scene optimization* and dense views, while sparse-view in-the-wild approaches are time-intensive. Feed-forward models [11, 39, 47] enable real-time reconstruction but are limited to fixed lighting and fail under dynamic conditions. In Tab. 1, we show the key characteristic differences between the previous methods. While these methods perform well on benchmarks like Photo-Tourism [33], they fail on more challenging datasets such as MegaScenes [37] (Fig. 2), which feature sparse views, diverse lighting, and heavy occlusions.

To address these limitations, we introduce **GenWild-Splat**, a generalizable model for fast feed-forward 3D scene reconstruction from sparse in-the-wild scenes without requiring per-scene optimization. To our knowledge, this is the first approach to integrate both appearance and occlusion modeling within a feed-forward 3D reconstruction paradigm. The key insight is that combining large-scale synthetic and real-world data enables the model to learn robust associations across diverse illumination conditions. Our framework uses VGGT's transformer to process sparse, unposed multi-view images into rich feature maps, which specialized heads decode into depth, camera parameters, and per-pixel Gaussians. The resulting set of attributes defines a canonical representation that captures a unified scene geometry disentangled from illumination. However, directly decoding these canonical Gaussians into a novel view via a differentiable rasterizer leads to multi-view inconsistencies.

To effectively map the canonical space to a desired target lighting, we introduce an appearance adapter that conditions on the target lighting and transforms the canonical Gaussian colors to the corresponding target lighting space. We parameterize lighting information using a light code estimated by a light encoder, which represents it in a compact latent space. To handle transient objects, we leverage a pre-trained segmentation network that identifies dynamic elements, such as people or cars. It produces explicit occlusion masks that guide our model to ignore transient regions during supervision, ensuring a clean and stable 3D scene. Ideally, one would train the model using multi-view images under varying illumination to render novel views under new lighting conditions. However, the absence of such multi-view, multi-lighting datasets makes direct supervised learning infeasible.

Table 1. Comparison of key characteristics across existing methods and our approach. Unlike optimization-based or feed-forward baselines, our method is fast, capable of sparse views, view-consistent, and generalizes to novel lighting conditions.

Method	In-the-Wild	Fast	View-Consis	Few-Views
Optimization-Based [16, 35]	✓	✗	✓	✗
Feed-Forward Based [11]	✗	✓	✗	✓
Ours	✓	✓	✓	✓

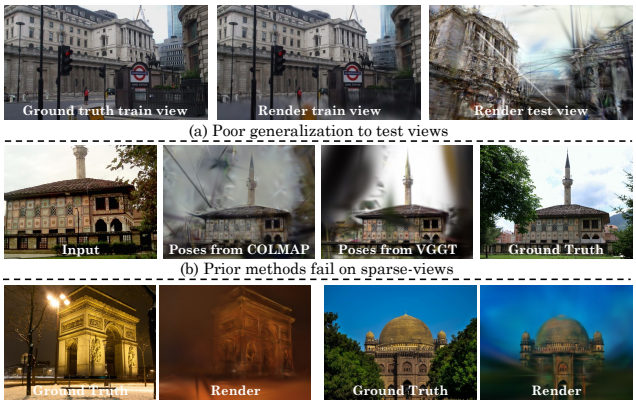

Figure 2. **Limitations of Prior Work.** Prior methods [16, 35] fail under sparse-view conditions. **(a) Overfitting:** Scene-specific optimization produces artifacts and geometric spikes with small camera perturbations. **(b) Camera dependency:** Methods rely on COLMAP for pose estimation, which fails under sparsity. Even with higher-quality transformer-based poses (e.g., VGGT), reconstructions exhibit severe artifacts and blurring. **(c) Limited appearance adaptation:** Test-time optimization cannot adapt to novel lighting, causing color bleeding and geometric distortions when target illumination differs from training conditions.

We train GenWildSplat without paired multi-view multi-illumination data, using unordered image collections. Each input image is mapped to a compact light code, and the appearance adapter conditions on this code and the canonical Gaussian colors to generate transformed colors, which are supervised via image reconstruction. Direct training on large-scale real-world data is unstable, as jointly learning geometry and illumination from sparse views is a highly ill-posed problem. To avoid collapse, we adopt a curriculum: first, learn appearance on synthetic data; and finally, add synthetic occlusions. This staged strategy enables stable optimization and strong generalization.

To evaluate generalization on unseen scenes, we benchmark GenWildSplat on both the Phototourism and the more challenging MegaScenes datasets. Our method consistently outperforms existing approaches [16, 35, 53] in reconstructing accurate scene geometry, modeling appearance under varying illumination, and effectively handling occlusions.

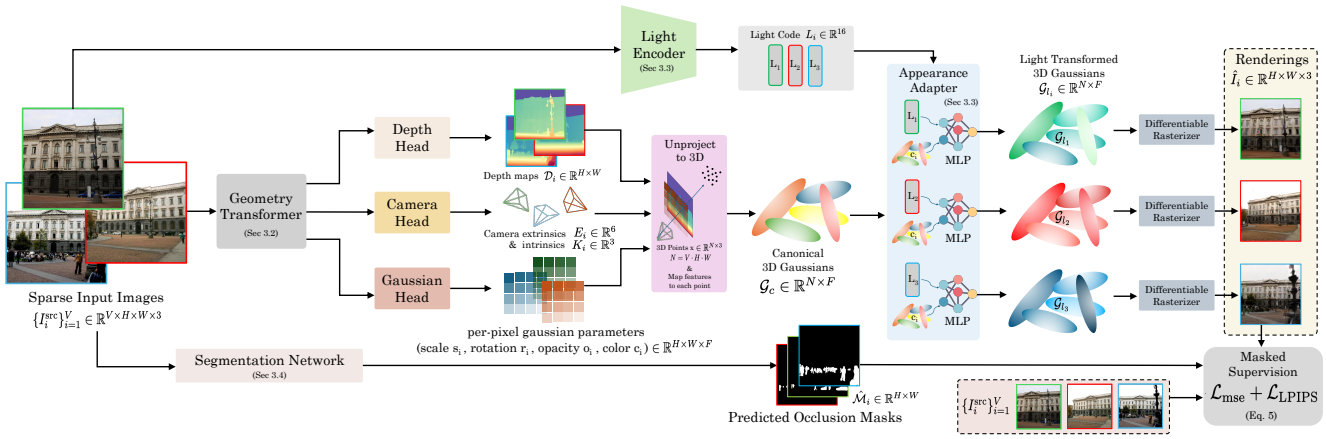

Figure 3. **Overview of GenWildSplat.** Given sparse, unposed images $\{I_i\}_{i=1}^V$ with appearance variations and transient objects, a geometry transformer extracts multi-view features $\mathbf{F}_i$ encoding semantic and geometric information. Specialized prediction heads process these features to output per-pixel depth $\mathbf{D}_i$, camera parameters $(\mathbf{K}_i, \mathbf{E}_i)$, and Gaussian attributes, which are unprojected into canonical 3D Gaussians $\mathbf{G}_c$. A light encoder $\mathcal{E}_{Light}$ extracts per-image lighting codes $\mathbf{L}_i = \mathcal{E}_{Light}(I_i)$. An MLP F_{light} modulates the canonical Gaussian colors using these codes: $\mathbf{G}_{\ell_i} = F_{\text{light}}(\mathbf{G}_c, \mathbf{L}_i)$. Each set of transformed Gaussians $\mathbf{G}_{\ell_i}$ is rasterized to reconstruct $\hat{I}_i$. A pre-trained segmentation network provides occlusion masks M_i to identify transient objects. Masked reconstruction loss focuses supervision on static content, enabling photorealistic, view-consistent reconstruction from sparse in-the-wild imagery.

Interestingly, GenWildSplat surpasses scene-specific methods [16, 35] that rely on per-image appearance optimization and test-time fine-tuning, demonstrating the strength of leveraging large-scale pre-trained priors for robust 3D understanding. Overall, GenWildSplat represents a step toward generalizable 3D reconstruction, offering a feed-forward, illumination- and occlusion-aware framework that scales to diverse, real-world environments for real-time 3D scene understanding.

2. Related Works

Optimization-based Novel View Synthesis (NVS) reconstructs 3D scenes from 2D images for novel viewpoint generation. *Gaussian Splatting (3DGS)* [15] represents scenes with explicit 3D Gaussian primitives, enabling real-time rasterization via a CUDA pipeline. Extensions improve depth and camera regularization for few-view settings [28, 46, 52, 57], but per-scene optimization is still required, limiting fast, test-time use.

Feed-forward Novel View Synthesis (NVS) predicts 3D Gaussians without scene-specific tuning, either assuming known poses (*pose-aware*) or estimating poses during inference (*pose-free*).

Pose-aware methods use calibrated poses and include: (1) direct 3D Gaussian predictors [3, 5, 6, 42, 47]; (2) transformer-based LRM decoders [9, 49, 54, 58]; and (3) latent feed-forward models [12]. These are fast but rely on accurate camera poses.

Pose-free methods jointly estimate poses and novel views, with DUSt3R [41] and MASt3R [17] predicting depth and fusing dense 3D. Subsequent works [23, 26, 36, 38–40, 50] extend this using transformer cascades for unified pose, trajectory, and geometry estimation, while latent approaches [10] self-supervise pose and view prediction. Despite strong performance, these methods degrade under lighting variation or dynamic distractors.

Novel View Synthesis in the Wild reconstructs 3D scenes from unconstrained photo collections, challenged by (*i*) varying illumination and (*ii*) transient objects.

Varying Appearance is handled via per-view latent embeddings [24, 51], CNN-conditioned features [43, 53], hash-grid fields [8], or hierarchical light decoupling [35]. Most require long test-time optimization (10+ hours). Diffusion-based strategies have also been explored to harmonize illumination across views [1] or to enable training-free multi-view consistent editing [2].

Occlusion Modeling addresses moving objects via robust regression [32], uncertainty features [16, 29], 2D occlusion masks [53], or per-image and per-Gaussian transient embeddings [35]. These methods remain slow, require intensive training, and result from a lack of 3D priors.

In contrast, our feed-forward approach directly processes sparse unposed images under varying lighting and dynamics, reconstructing 3D scenes with controllable appearance and view-consistent rendering in 3 seconds, without per-scene optimization.

3. Method

3.1. Preliminary: AnySplat

Our method builds upon AnySplat [11], a feed-forward framework that reconstructs 3D scenes as Gaussian primitives from multiple input images in a single pass.

Figure 4. **Curriculum Learning.** Training proceeds in three stages. **Stage I:** Single scene with illumination variation. In this stage, the model learns to disentangle lighting from geometry. **Stage II:** Multiple scenes: the model then learns geometric and appearance priors across diverse environments. **Stage III:** Synthetic occlusions: the network learns to handle transient objects and multi-view inconsistencies. Despite training only on synthetic data, the model generalizes to real-world appearance variations and occlusions.

Architecture. AnySplat takes unposed images and processes them through a VGGT [39] transformer backbone to extract multi-view features. Three prediction heads then output: (a) depth maps for each view, (b) camera poses and (c) 3D Gaussian properties including position, color, shape, and opacity. To reduce redundancy from per-pixel Gaussian prediction, AnySplat voxelizes the scene, assigns confidence scores, and merges overlapping Gaussians within each voxel to form a compact 3D representation.

Training. AnySplat trains without ground-truth 3D data. Instead, it uses VGGT's pretrained model to generate pseudo-labels for depth and camera poses. The predicted Gaussians are rendered back to 2D and supervised against the input views, ensuring the 3D representation remains consistent with the observed images.

3.2. Problem Formulation and Overview

Given unposed input images $\mathcal{I} = \{I_1, I_2, \dots, I_N\}$ captured under varying illumination with transient objects, we reconstruct a 3D scene that renders novel views under different appearance conditions while handling occlusions.

Our model predicts 3D Gaussians conditioned on target appearance $\mathbf{L}$ as $\boldsymbol{\mathcal{G}_l} = f_\theta(\mathcal{I}, \mathbf{L})$, where each Gaussian $g_L \in \boldsymbol{\mathcal{G}_l}$ is parameterized by:

$$g_l = \{\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{c}_L\},$$

with position $\boldsymbol{\mu} \in \mathbb{R}^3$, opacity $\boldsymbol{\sigma} \in \mathbb{R}^+$, rotation $\boldsymbol{r} \in \mathbb{R}^4$, scale $\boldsymbol{s} \in \mathbb{R}^3$, and appearance-dependent spherical harmonic (SH) coefficients $\boldsymbol{c}_L \in \mathbb{R}^{75}$.

Architecture. A VGGT transformer backbone ϕ_θ extracts multi-view features $\boldsymbol{F} = \phi_\theta(\mathcal{I})$. Similar to Anysplat, three prediction heads process these features:

$$\boldsymbol{D} = h_D(\boldsymbol{F}),\ (\boldsymbol{K}, \boldsymbol{E}) = h_C(\boldsymbol{F}),\ (\boldsymbol{s}, \boldsymbol{r}, \boldsymbol{\sigma}, \boldsymbol{c}) = h_{\text{gauss}}(\boldsymbol{F}),$$

where h_D predicts per-view depth maps $\boldsymbol{D}$, h_C estimates camera intrinsics $\boldsymbol{K}$ and extrinsics $\boldsymbol{E}$, and h_{gauss} outputs appearance-independent Gaussian properties and canonical SH coefficients $\boldsymbol{c} \in \mathbb{R}^{75}$. An appearance adapter ψ_θ modulates the canonical colors for target appearance:

$$\boldsymbol{c}_{\boldsymbol{L}_i} = \psi_\theta(\boldsymbol{c}, \mathbf{L}_i), \tag{1}$$

where $\mathbf{L}_i \in \mathbb{R}^d$ is a learned appearance embedding.

Training. Gaussians $\boldsymbol{\mathcal{G}_l}$ are rasterized via diff. splatting:

$$\hat{I}_j = \mathcal{R}(\boldsymbol{\mathcal{G}_l}, \boldsymbol{K}_j, \boldsymbol{E}_j), \tag{2}$$

and trained end-to-end with reconstruction loss. Though trained only on input views, the model generalizes to novel appearance conditions. We adopt a curriculum training strategy to sequentially refine geometry, appearance, and occlusion modeling for stable convergence. We describe our methodology in Fig. 3

3.3. Appearance Modelling: Appearance Adapter

Existing methods like WildGaussians [16] and NexusSplats [35] model appearance using randomly initialized embeddings jointly optimized with geometry during training. At test time, these methods require optimizing a new embedding for each novel view or lighting condition, precluding feed-forward inference. We instead predict all scene parameters, including appearance, in a single forward pass.

Our *Appearance Adapter* transforms Gaussian colors to match a target lighting condition. A 2D CNN-based encoder $\mathcal{E}_{Light}$ extracts per-view light codes, which an MLP F_{light} uses to modulate the Gaussian colors $\mathcal{G}_c = [\mathbf{c}_1, \dots, \mathbf{c}_N]^\top \in \mathbb{R}^{N \times 75}$:

$$\mathbf{L}_i = \mathcal{E}_{Light}(I^{(i)}), \quad i = 1, \dots, V. \tag{3}$$

$$\mathcal{G}_{l_i} = F_{\text{light}}(\mathcal{G}_c, \mathbf{L}_i), \quad i = 1, \dots, V, \tag{4}$$

where $\mathcal{G}_{l_i} = [\tilde{\mathbf{c}}_{i,1}, \dots, \tilde{\mathbf{c}}_{i,N}]^\top$ are the transformed colors under view i's lighting. Each set of transformed Gaussians is independently rasterized to reconstruct its corresponding input view, enabling self-supervised training without test-time optimization.

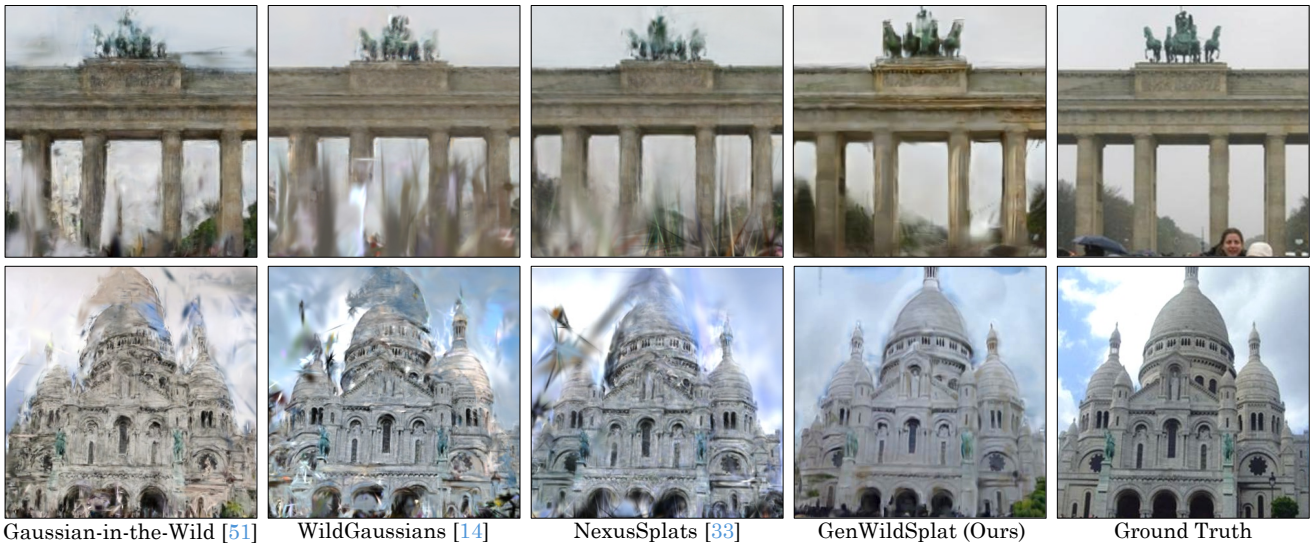

Figure 5. **Comparison on the Photo-Tourism dataset against optimization-based methods.** Optimization-based methods trained from scratch often struggle to accurately reconstruct scenes from sparse views, even when test-time optimization is applied. In contrast, our feedforward approach efficiently generates plausible geometry and controllable appearance for complex scenes. As shown in Fig. 2, replacing COLMAP poses with VGGT poses improves their performance; thus, we adopt this modification across all evaluations, thereby solely benefiting the baseline's performance.

3.4. Occlusion Modelling

Transient objects (such as people and vehicles) cause floating artifacts and unstable gradients when treated as static geometry. Prior work [53] uses internally predicted visibility maps or uncertainty estimates [16, 27], which can collapse during unsupervised training by down-weighting difficult regions. This incorrectly suppresses static structures, such as trees, that appear in sparse views (Fig. 2).

We instead use a pre-trained semantic segmentation network to detect common transient classes (person, car, bus, truck). These predictions yield a binary mask $S \in \{0,1\}^{H\times W}$ where $S(p) = 1$ indicates transients. We apply visibility weighting $M = 1 - S$ directly to images: $I_{\text{m}} = I \odot M$ and $\hat{I}_{\text{m}} = \hat{I} \odot M$, focusing on static regions:

$$\mathcal{L} = \text{MSE}(I_{\text{m}}, \hat{I}_{\text{m}}) + \lambda \cdot \text{Percep}(I_{\text{m}}, \hat{I}_{\text{m}}) \tag{5}$$

where $\hat{I}$ is the rendered image, I_{gt} the ground truth, and $\odot$ denotes elementwise multiplication. Using an external segmentation prior in this way prevents the model from "explaining away" transient content by collapsing its own visibility estimate, stabilizes gradients in dynamic regions, and preserves the static structure during training.

3.5. Curriculum Learning for Large-Scale Training

Feed-forward reconstruction on unconstrained imagery requires training on large, diverse datasets. Direct training on data with appearance variation and transient objects is unstable, as learning geometry, lighting, and occlusion jointly is difficult. Training only on curated datasets, however, fails to build priors for in-the-wild generalization. We use curriculum learning to break the task into progressive stages (Fig. 4), thereby improving convergence and reconstruction quality compared to end-to-end training.

Stage 1: Lighting (Appearance). Train on a single synthetic scene with illumination variation but no transients, learning lighting or broadly appearance representation without geometric or occlusion confounds. Empirically, we found that this simplified the appearance decomposition for subsequent training without collapsing.

Stage 2: Multi-scene generalization. Introduce additional synthetic scenes to improve appearance and geometry modeling across diverse environments.

Stage 3: Occlusion handling. Add synthetic transients where we have access to ground-truth masks for supervision. We then train the model to predict these occlusion masks alongside geometry and appearance, disentangling transients from static content.

Despite being trained only on synthetic occlusions and appearance variations, our method generalizes well to real-world sparse-view scenes (Fig. 5, Fig. 6).

3.6. Training Framework

For each input image, the network predicts scene geometry (per-Gaussian parameters and depth), while the light encoder extracts a compact light code that represents the image's illumination. The appearance adapter conditions on this code and the canonical Gaussian colors to produce transformed colors, which are rasterized and compared to the original image. Though the method is not trained to

Figure 6. **Comparison on the MegaScenes dataset against optimization-based methods.** The MegaScenes dataset poses significant challenges for 3D reconstruction due to wide variations in viewpoints and lighting. Prior SOTA methods often fail, producing artifacts such as noisy ground (row 1), geometric distortions and inconsistencies when rendering novel views (row 2), and spiky/blurred skies (row 3). GenWildSplat, in contrast, generates clean and consistent renderings across diverse scenes, demonstrating robust performance even on these highly challenging in-the-wild settings.

render novel views or lighting, our method generalizes well to unseen views (Fig. 6). The network learns stable lighting representations that enable transferring illumination appearance from one scene to another (Fig. 8), a capability absent in prior in-the-wild methods [16, 35, 53].

4. Experiments

4.1. Implementation Details

GenWildSplat uses a 24-layer transformer with alternating frame and global attention. The depth, camera, and Gaussian heads adopt a DPT-based architecture that fuses multi-scale features to predict per-pixel depth, Gaussian attributes, and camera parameters. The light encoder follows [55], producing 16-dimensional lighting vectors. An MLP expands these to 75 dimensions and modulates the per-Gaussian SH coefficients accordingly. For occlusion detection, we use YOLOv8 Segmentation [13] to classify common COCO categories (person, car, dog, etc.) and merge them into a binary transient mask. The model, initialized from AnySplat pre-trained weights, uses a perceptual loss weight of $\lambda = 0.05$ and is trained via curriculum learning for 40K iterations (Stage 1: 10K, Stage 2: 10K, Stage 3: 20K) over 2 days on a single RTX A6000. Please refer to supplementary for more details.

4.2. Datasets

Training. We train on 700+ outdoor scenes from DL3DV [22], augmented with synthetic lighting and occlusions to mimic in-the-wild variability. Illumination diversity is produced using DiffusionRenderer [20] via offline unconditioned relighting (30 minutes per scene). Transient occluders are generated by compositing COCO segments [21] (e.g., people, cars) at random locations, providing exact ground-truth masks and are applied on-the-fly.

Evaluation. We evaluate on PhotoTourism [33] using 6 input views across 3 scenes. To assess generalization, we further curate 20 challenging MegaScenes with strong lighting variation, occlusions, and viewpoint sparsity, selecting scenes with fewer than 20 registered images. This avoids artificial subsampling and provides a more realistic sparse-view benchmark for the future. Refer to the supplementary for visualizations of these sparse-view scenes.

Figure 7. **Comparison on the MegaScenes dataset against feed-forward based methods.** Existing feed-forward 3D Gaussian Splatting methods cannot handle unconstrained inputs, so we construct baselines using style transfer and DiffusionRenderer to address appearance variations. The DiffusionRenderer+AnySplat baseline integrates AnySplat with DiffusionRenderer, which uses environment maps from DiffusionLight-Turbo. Style transfer [45] often introduces artifacts and color bleeding, while DiffusionRenderer [20] produces unrealistic outdoor relighting (row 2 shows a dimmed, non-photorealistic "night"). These per-image methods suffer from multi-view inconsistency, whereas GenWildSplat modulates appearance in 3D, yielding photorealistic, view-consistent results.

Figure 8. **Cross-scene appearance transfer.** Our method disentangles appearance from geometry, allowing adaptation of illumination from *different* scenes, something prior methods [16, 35] cannot do as they jointly optimize view and appearance.

4.3. Baselines

Recent in-the-wild reconstruction methods, such as GS-W [53], WildGaussians [16], and NexusSplats [35], require per-scene and test-time optimization, making a direct comparison with our feed-forward approach infeasible. Methods like SparseGS-W [19] and MS-GS [18] lack public implementations. We therefore define feed-forward baselines:

AnySplat [11] provides real-time reconstruction but does not model appearance variation.

StyleTransfer-AnySplat extends AnySplat with CCPL [45] for style adaptation, though this adjusts artistic style rather than realistic illumination.

DiffusionRenderer+AnySplat integrates AnySplat with DiffusionRenderer [20], which models lighting using environment maps from DiffusionLight-Turbo [7].

All baselines use Stable Diffusion [30] for mask-based inpainting to handle occlusions. We use AnySplat as our primary baseline, since GenWildSplat builds upon it; however, our modular approach could also extend to other feed-forward methods, such as MVSplat [5] or PixelSplat [3].

Table 2. **Quantitative comparison against baselines.** We compare against in-the-wild baselines on MegaScenes under sparse-view settings with varying input views, where the best scores and second best scores are highlighted with respective colors.

Method	Gen.	Time	MegaScenes (3-View)			MegaScenes (6-View)		
			PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
GS-W	✗	5 hrs	11.60	0.285	0.623	12.01	0.312	0.552
WildGaussians	✗	8 hrs	12.73	0.316	0.599	13.29	0.373	0.532
NexusSplats	✗	2.4 hrs	13.17	0.335	0.552	13.92	0.397	0.518
GenWildSplat (Ours)	✓	3 secs	14.43	0.402	0.496	15.84	0.440	0.407

Table 3. **Quantitative comparison against feed-forward methods.** We compare our method against the feed-forward baselines on the sparse-view setting on the MegaScenes dataset.

Method	View Consistent	PSNR	SSIM	LPIPS
Vanilla AnySplat	✗	12.65	0.311	0.412
2D Baseline + AnySplat	✗	12.90	0.281	0.486
DiffusionRenderer + AnySplat	✗	13.59	0.309	0.444
GenWildSplat (Ours)	✓	15.84	0.440	0.407

4.4. Comparison on the PhotoTourism Dataset

We evaluate GenWildSplat against state-of-the-art in-the-wild baselines [16, 35, 53] on sparse-view PhotoTourism

Table 4. **Ablation study** evaluated on the MegaScenes Dataset.

Model Variant	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Appearance adapter	13.76	0.391	0.405
w/o Occlusion handling	15.14	0.405	0.513
w/o Curriculum learning	11.72	0.318	0.438
Full model (Ours)	15.84	0.440	0.407

(Fig. 5, Tab. 2). As Fig. 2 shows, COLMAP poses often degrade baseline performance under sparse views, so we use VGGT poses for fair comparison. Despite no scene-specific training, our feed-forward model surpasses optimization-based methods, producing more realistic renderings from sparse inputs. This stems from our appearance adapter, which transfers priors learned via curriculum training, enabling inference in just 3 seconds.

4.5. Comparison on the MegaScenes Dataset

We further evaluate GenWildSplat on the challenging MegaScenes dataset (Fig. 6, Tab. 3). Prior methods, trained from scratch without learned priors, exhibit severe artifacts and distortions, such as noisy ground regions(row 1), poor generalisation to novel views(row 2), and resulting in blurred skies(row 3). GenWildSplat produces clean, consistent renderings across diverse scenes, demonstrating its robustness even on very challenging in-the-wild datasets.

For a fair comparison, we benchmark against a few AnySplat variants (Fig. 7). 2D style transfer [45] often introduces artifacts and color bleeding, while DiffusionRenderer [20], relying on estimated environment maps, produces unrealistic outdoor relighting (e.g., row 2 shows a dimmed but non-photorealistic "night"). These per-image methods suffer from multi-view inconsistency, unlike GenWildSplat, which modulates appearance directly in 3D for photorealistic, view-consistent results.

4.6. Results with Lighting from Different Scene

Unlike prior in-the-wild methods that jointly optimize lighting and geometry and require a target lighting image from the same scene, GenWildSplat disentangles appearance from geometry, enabling cross-scene illumination transfer. As shown in Fig. 8, it produces photorealistic, view-consistent results while preserving spatial and structural consistency, demonstrating robust appearance control.

4.7. Ablation Study & Analysis

To evaluate the contribution of each component in our method, we perform an ablation study shown in Fig. 9 and Tab. 4. Removing the appearance adapter prevents the model from capturing appearance variations, resulting in a fixed, single appearance. Disabling occlusion handling prevents the removal of transient objects, such as people on the stairs. Without the proposed curriculum-based training, the Gaussian colors collapse, as the model struggles to learn

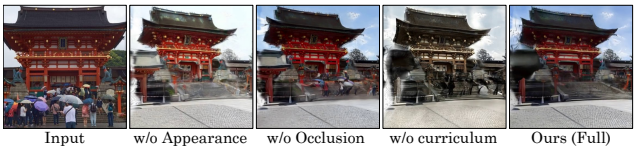

Figure 9. **Ablation Study.** Removing the appearance adapter, occlusion handling, or curriculum causes major failures: fixed appearance, baked-in transient objects, or color collapse. With all components enabled, GenWildSplat produces clean, consistent 3D reconstructions.

Figure 10. **Limitations.** (a) missing geometry in sparsely observed regions, (b) artifacts and double geometry for test views distant from training views, (c) degraded performance in indoor environments with imperfect occlusion masks, and (d) absence of shadow modeling and realistic relighting.

geometry, appearance, and occlusions simultaneously. With all components enabled, GenWildSplat models both appearance and occlusions, producing view-consistent renderings.

5. Discussions

Limitations. GenWildSplat, while effective under sparse, in-the-wild image collections, has several limitations. First, sparse viewpoints naturally leave unseen regions, leading to incomplete geometry in areas not covered by the input images. Second, when test views lie far outside the training distribution, the model may produce artifacts or double-layered geometry due to limited viewpoint generalization. Third, indoor scenes are still hard: if the occlusion mask fails to accurately capture objects or depth discontinuities, the resulting masks degrade the reconstruction quality. Finally, the method does not model cast shadows or support realistic relighting, limiting its applicability to tasks that require physically consistent illumination.

Conclusion. We present GenWildSplat, a generalizable, feed-forward Gaussian-splatting framework that reconstructs 3D scenes from sparse, unconstrained photo collections in under 3 seconds. The key to our success is the appearance adapter, which directly modulates Gaussian colors in 3D, and a robust occlusion handling mechanism, producing view-consistent, photorealistic renderings. GenWildSplat moves the needle towards real-time, controllable, relightable 3D scenes from sparse internet imagery.

References

[1] Hadi Alzayer, Philipp Henzler, Jonathan T Barron, Jia-Bin Huang, Pratul P Srinivasan, and Dor Verbin. Generative multiview relighting for 3d reconstruction under extreme illumination variation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10933–10942, 2025. 3

[2] Hadi Alzayer, Yunzhi Zhang, Chen Geng, Jia-Bin Huang, and Jiajun Wu. Coupled diffusion sampling for training-free multi-view image editing. *arXiv preprint arXiv:2510.14981*, 2025. 3

[3] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024. 3, 7

[4] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *CVPR*, 2022. 2

[5] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, 2024. 3, 7

[6] Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *NeurIPS*, 2024. 3

[7] Worameth Chinchuthakun, Pakkapon Phongthawee, Amit Raj, Varun Jampani, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight-turbo: Accelerated light probes for free via single-pass chrome ball inpainting. *arXiv preprint arXiv:2507.01305*, 2025. 7

[8] Hiba Dahmani, Moussab Bennehar, Nathan Piasco, Luis Roldao, and Dzmitry Tsishkou. Swag: Splatting in the wild images with appearance-conditioned gaussians. In *ECCV*, 2024. 2, 3

[9] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *ICLR*, 2023. 3

[10] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *ICCV*, 2025. 3

[11] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. In *ACM SIGGRAPH Asia*, 2025. 2, 3, 7

[12] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *ICLR*, 2024. 3

[13] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 6

[14] Joanna Kaleta, Kacper Kania, Tomasz Trzciński, and Marek Kowalski. Lumigauss: Relightable gaussian splatting in the wild. In *WACV*, 2025. 2

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[16] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. In *NeurIPS*, 2024. 2, 3, 4, 5, 6, 7

[17] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 3

[18] Deming Li, Kaiwen Jiang, Yutao Tang, Ravi Ramamoorthi, Rama Chellappa, and Cheng Peng. Ms-gs: Multi-appearance sparse-view 3d gaussian splatting in the wild. *arXiv preprint arXiv:2509.15548*, 2025. 7

[19] Yiqing Li, Xuan Wang, Jiawei Wu, Yikun Ma, and Zhi Jin. Sparsegs-w: Sparse-view 3d gaussian splatting in the wild with generative priors. *arXiv preprint arXiv:2503.19452*, 2025. 7

[20] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *CVPR*, 2025. 6, 7, 8

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[22] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. 6

[23] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *CVPR*, 2025. 3

[24] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 3

[25] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[26] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *CVPR*, 2025. 3

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 5

[28] Avinash Paliwal, Wei Ye, Jinhui Xiong, Dmytro Kotovenko, Rakesh Ranjan, Vikas Chandra, and Nima Khademi Kalantari. Coherentgs: Sparse novel view synthesis with coherent 3d gaussians. In *ECCV*, 2024. 3

[29] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting

uncertainty for distractor-free nerfs in the wild. In *CVPR*, 2024. 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 7

[31] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *ECCV*, 2022. 2

[32] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *CVPR*, 2023. 3

[33] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*. 2006. 2, 6

[34] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 conference proceedings*, 2022. 2

[35] Yuzhou Tang, Dejun Xu, Yongjie Hou, Zhenzhong Wang, and Min Jiang. Nexussplats: Efficient 3d gaussian splatting in the wild. *arXiv preprint arXiv:2411.14514*, 2024. 2, 3, 4, 6, 7

[36] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *CVPR*, 2025. 3

[37] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024. 2

[38] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, 2025. 3

[39] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 2, 4

[40] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 3

[41] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 3

[42] Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. *NeurIPS*, 37, 2024. 3

[43] Yuze Wang, Junyi Wang, and Yue Qi. We-gs: An in-the-wild efficient 3d gaussian representation for unconstrained photo collections. *arXiv preprint arXiv:2406.02407*, 2024. 2, 3

[44] Yuze Wang, Junyi Wang, Ruicheng Gao, Yansong Qu, Wantong Duan, Shuo Yang, and Yue Qi. Look at the sky: Sky-aware efficient 3d gaussian splatting in the wild. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[45] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: Contrastive coherence preserving loss for versatile style transfer. In *ECCV*, 2022. 7, 8

[46] Haolin Xiong, Sairisheek Muttukuru, Hanyuan Xiao, Rishi Upadhyay, Pradyumna Chari, Yajie Zhao, and Achuta Kadambi. Sparsegs: Sparse view synthesis using 3d gaussian splatting. In *2025 International Conference on 3D Vision (3DV)*, 2025. 3

[47] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *CVPR*, 2025. 2, 3

[48] Jiacong Xu, Yiqun Mei, and Vishal Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. In *NeurIPS*, 2024. 2

[49] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *ECCV*, 2024. 3

[50] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *CVPR*, 2025. 3

[51] Yifan Yang, Shuhai Zhang, Zixiong Huang, Yubing Zhang, and Mingkui Tan. Cross-ray neural radiance fields for novel-view synthesis from unconstrained image collections. In *ICCV*, 2023. 3

[52] Ruihong Yin, Vladimir Yugay, Yue Li, Sezer Karaoglu, and Theo Gevers. Fewviewgs: Gaussian splatting with few view matching and multi-stage training. *NeurIPS*, 2024. 3

[53] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *ECCV*, 2024. 2, 3, 5, 6, 7

[54] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *ECCV*, 2024. 3

[55] Xiao Zhang, William Gao, Seemandhar Jain, Michael Maire, David Forsyth, and Anand Bhattad. Latent intrinsics emerge from training to relight. In *NeurIPS*, 2024. 6

[56] Linglong Zhou, Guoxin Wu, Yunbo Zuo, Xuanyu Chen, and Hongle Hu. A comprehensive review of vision-based 3d reconstruction methods. *Sensors*, 24(7):2314, 2024. 2

[57] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *ECCV*, 2024. 3

[58] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *ICCV*, 2025. 3