

FreeSplatter: Pose-free Gaussian Splatting for Sparse-view 3D Reconstruction

Jiale Xu¹ Shenghua Gao² Ying Shan¹

¹ARC Lab, Tencent PCG ²The University of Hong Kong

<https://bluestyle97.github.io/projects/freesplatter/>

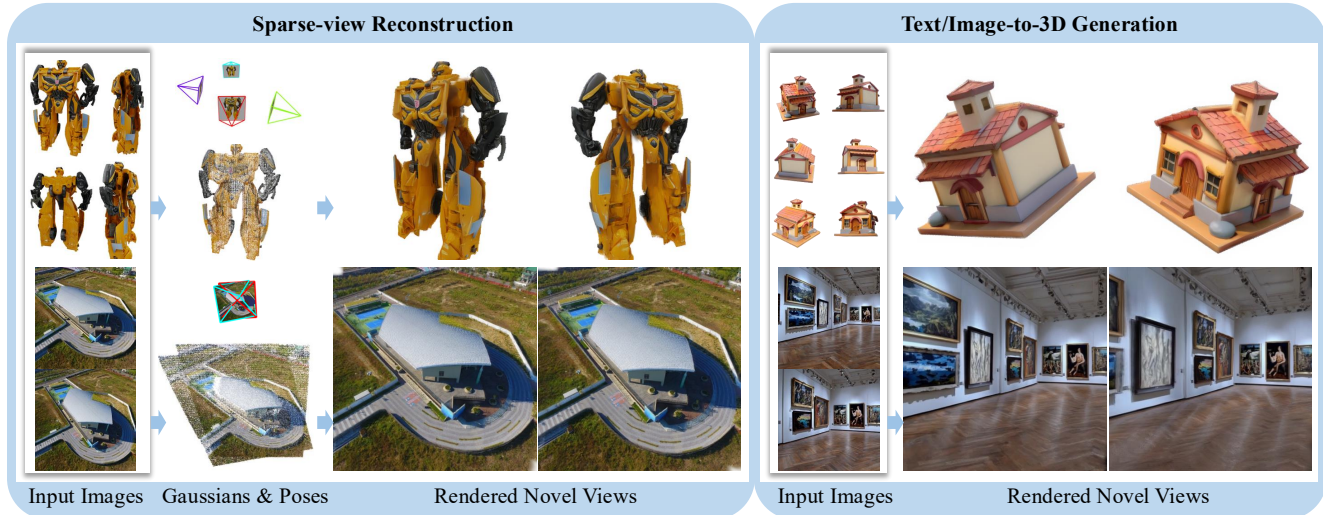


Figure 1. FreeSplatter reconstructs high-fidelity 3D Gaussians and estimates accurate camera poses from uncalibrated sparse-view images in a feed-forward manner, handling both object-centric (1st row) and scene-level (2nd row) scenarios effectively. It can also seamlessly process synthetic multi-view images from diffusion models, enabling efficient and high-quality text/image-to-3D content creation.

Abstract

Sparse-view reconstruction models typically require precise camera poses, yet obtaining these parameters from sparse-view images remains challenging. We introduce **FreeSplatter**, a scalable feed-forward framework that generates high-quality 3D Gaussians from **uncalibrated** sparse-view images while estimating camera parameters within seconds. Our approach employs a streamlined transformer architecture where self-attention blocks facilitate information exchange among multi-view image tokens, decoding them into pixel-aligned 3D Gaussian primitives within a unified reference frame. This representation enables both high-fidelity 3D modeling and efficient camera parameter estimation using off-the-shelf solvers. We develop two specialized variants—for **object-centric** and **scene-level** reconstruction—trained on comprehensive datasets. Remarkably, FreeSplatter outperforms several pose-dependent Large Reconstruction Models (LRMs) by a notable margin while achieving comparable or even better pose estimation accuracy compared to state-of-the-art pose-free reconstruction approach MAST3R in challeng-

ing benchmarks. Beyond technical benchmarks, FreeSplatter streamlines text/image-to-3D content creation pipelines, eliminating the complexity of camera pose management while delivering exceptional visual fidelity.

1. Introduction

Recent breakthroughs in neural scene representation and differentiable rendering, e.g., Neural Radiance Fields (NeRF)[35] and Gaussian Splatting (GS)[27], have demonstrated exceptional multi-view reconstruction quality for densely-captured images with calibrated camera poses through per-scene optimization. However, these approaches fail in sparse-view scenarios where traditional camera calibration techniques like Structure-from-Motion (SfM)[40] struggle due to insufficient image overlaps. While generalizable reconstruction models[5, 23, 57] address sparse-view reconstruction using learned priors in a feed-forward manner, they still require accurate camera parameters, sidestepping a fundamental challenge in real-world applications. Liberating sparse-view reconstruction from known camera poses remains a critical frontier.

Previous pose-free reconstruction efforts include PF-LRM [49] and LEAP [26], which map multi-view image tokens to NeRF representations using transformers. Despite their pioneering contributions, their approaches suffer from inefficient volume rendering and limited resolution, hampering training efficiency and scalability to complex scenes. Moreover, inferring camera poses from their implicit representations requires additional specialized components, introducing extra complexity. DUST3R [51] presents an alternative paradigm for joint 3D reconstruction and pose estimation through direct point regression, enabling efficient camera pose recovery with PnP solvers [19, 20] and demonstrating impressive zero-shot capabilities.

However, point clouds’ inherent sparsity limits their utility for downstream applications like novel view synthesis. In contrast, 3D Gaussian Splats (3DGS) can encode high-fidelity radiance fields while enabling efficient rendering by augmenting point clouds with additional attributes. This raises the question: can we directly predict “Gaussian maps” from multi-view images to achieve both high-quality 3D modeling and instant camera pose estimation?

We introduce FreeSplatter, a feed-forward reconstruction framework that jointly predicts pixel-wise Gaussians from uncalibrated sparse-view images and estimates their camera parameters. At its core is a scalable streamlined transformer that maps multi-view image tokens into pixel-aligned Gaussian maps using simple self-attention layers—requiring no camera poses, intrinsics, or post-alignment. These Gaussian maps enable both high-fidelity scene representation and ultra-fast camera parameter estimation using off-the-shelf solvers [19, 20, 36].

Leveraging the training and rendering efficiency of 3D Gaussians, we extend our approach to complex scene-level reconstruction by training two variants: FreeSplatter-O for object-centric reconstruction (trained on Objaverse [9]) and FreeSplatter-S for scene-level reconstruction (trained on mixed datasets [37, 61, 62]). Both share a common architecture with task-specific adjustments. Our extensive experiments demonstrate FreeSplatter’s superiority over existing methods in both reconstruction quality and pose estimation accuracy. Notably, FreeSplatter-O significantly outperforms several existing *pose-dependent* large reconstruction models, while FreeSplatter-S achieves comparable or better pose estimation accuracy than state-of-the-art MAST3R [28] across challenging benchmarks. We further demonstrate FreeSplatter’s potential for enhancing 3D content creation pipelines through integration with multi-view diffusion models.

2. Related Work

Large Reconstruction Models. Large-scale 3D object datasets [9, 10] have enabled training of highly generalizable models for open-category image-to-3D reconstruc-

tion. Large Reconstruction Models (LRMs) [23, 29, 59] employ scalable feed-forward transformer architectures to map sparse-view image tokens into 3D triplane NeRF representations [4, 35], supervised with multi-view rendering losses. Recent advances have explored alternative representations including meshes [52, 54, 57] and 3D Gaussians [44, 58, 65] for real-time rendering, more efficient network architectures [3, 6, 30, 63, 66], enhanced texture quality [1, 41, 60], and explicit 3D supervision for improved geometry [33]. Despite their impressive reconstruction quality and generalization capabilities, LRMs require *posed* images as input and are highly sensitive to pose accuracy, significantly limiting their practical application scenarios.

Pose-free Reconstruction. Classical pose-free reconstruction algorithms like Structure from Motion (SfM) [20, 40, 45] first establish pixel correspondences across multiple views, then perform 3D point triangulation and bundle adjustment to jointly optimize 3D coordinates and camera parameters. Recent improvements to SfM leverage learning-based feature descriptors [11, 14, 38, 50], image matchers [15, 16, 32, 39], and differentiable bundle adjustment [31, 47, 53]. While effective with sufficient image overlaps, SfM-based methods struggle with sparse views where correspondence matching becomes challenging. Learning-based methods [17, 22, 25, 26] address this by utilizing learned data priors to recover 3D geometry from input views. PF-LRM [49] extends the LRM framework by predicting per-view coarse point clouds for camera pose estimation with a differentiable PnP solver [7]. DUST3R [51] introduces a novel approach to Multi-view Stereo (MVS) by framing it as a pointmap-regression problem, with subsequent works enhancing its local representation capabilities [28] and reconstruction efficiency [46].

Generalizable Gaussian Splatting. Compared to the implicit MLP-based representation of NeRF [35], 3D Gaussian Splatting (3DGS) [24, 27] explicitly represents scenes as point clouds with additional attributes, achieving an balance between high-fidelity rendering and real-time performance. However, traditional 3DGS requires per-scene optimization with densely-captured images and SfM-generated sparse point clouds for initialization. Recent research [5, 8, 34, 43, 55] has explored feed-forward models for sparse-view Gaussian reconstruction by leveraging large-scale datasets and scalable architectures. These approaches typically assume access to accurate camera poses and employ 3D-to-2D geometric projection for feature aggregation, using techniques like epipolar lines [5] or plane-swept cost volumes [8, 34]. InstantSplat [18] and Splatt3R [42] leverage DUST3R/MASt3R’s pose-free reconstruction capabilities—the former initializes Gaussian positions using DUST3R point clouds before optimizing other Gaussian parameters, while the latter trains a Gaussian head on a frozen MAST3R model. Despite impressive results, their recon-

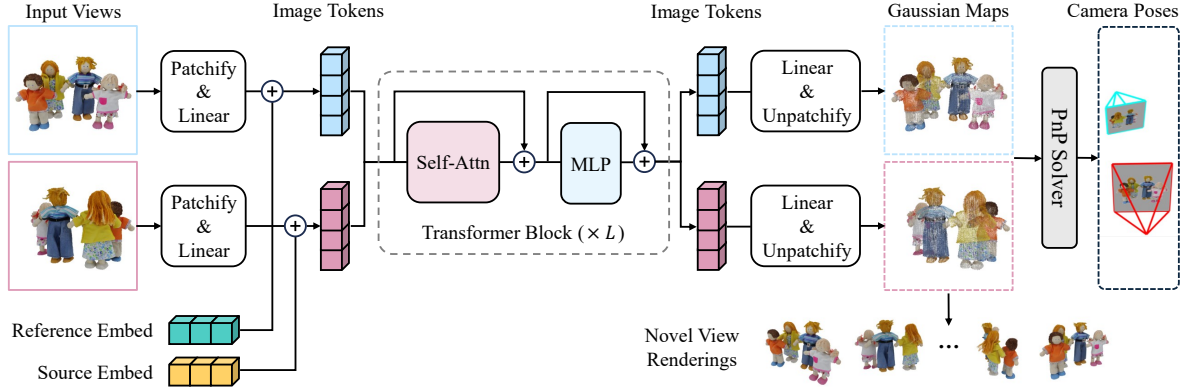


Figure 2. **FreeSplatter Pipeline.** Given N uncalibrated input views without any known camera extrinsics or intrinsics, we first patchify each image into tokens and feed these tokens into a sequence of self-attention blocks, enabling information exchange across multiple views. The resulting tokens are then decoded into N Gaussian maps, which allow us to render novel views and simultaneously recover the camera focal length f and poses using simple iterative solvers.

struction quality remains heavily dependent on the quality of the initial point clouds generated by DUST3R.

3. Method

Given N input images $\{I^n \mid n = 1, \dots, N\}$ without known camera parameters, FreeSplatter performs joint scene reconstruction and camera parameter estimation. The pipeline is formulated as:

$$G, P^1, \dots, P^N, f = \text{FreeSplatter}(I^1, \dots, I^N), \quad (1)$$

where $G = \{G^n \mid n = 1, \dots, N\}$ represents the unified set of reconstructed Gaussian maps, P^n denotes the estimated camera pose for I^n , and f represents the shared focal length across views (reasonable in most scenarios).

3.1. Preliminary

3D Gaussian Splatting (3DGS) [27] represents a scene as a set of 3D Gaussian primitives. Each primitive is parameterized by location $\mu_k \in \mathbb{R}^3$, rotation quaternion $r_k \in \mathbb{R}^4$, scale $s_k \in \mathbb{R}^3$, opacity $o_k \in \mathbb{R}$, and Spherical Harmonic (SH) coefficients $c_k \in \mathbb{R}^{3 \times d^2}$ for computing view-dependent color (d denoting the degree of SH). This representation parameterizes scene radiance fields through explicit point clouds, enabling efficient novel view synthesis via rasterization. Compared to NeRF’s computationally intensive volume rendering, 3DGS achieves comparable visual quality with significantly reduced computational and memory requirements.

3.2. Model Architecture

As Figure 2 shows, FreeSplatter adopts a transformer architecture inspired by GS-LRM [65]. For input images $\{I^n \mid n = 1, \dots, N\}$, the model patchifies them into tokens $\{e^{n,m} \mid n = 1, \dots, N, m = 1, \dots, M\}$ (M denotes patch number per image), processes them through self-attention

blocks for multi-view information exchange, and decodes them into N Gaussian maps $\{G^n \mid n = 1, \dots, N\}$. These maps enable novel view synthesis and camera parameter recovery through iterative optimization.

Image Tokenization. The model processes N input images $\{I^n \in \mathbb{R}^{H \times W \times 3} \mid n = 1, \dots, N\}$ using ViT-style [12] tokenization: images are divided into $p \times p$ patches ($p = 8$), flattened to $p^2 \cdot 3$ dimensional vectors, and projected to d -dimensional tokens via a linear layer. Each token $e^{n,m}$ is enhanced with position and view embeddings:

$$e^{n,m} = e^{n,m} + e_{\text{pos}}^m + e_{\text{view}}^n, \quad (2)$$

where e_{pos}^m encodes patch position and e_{view}^n distinguishes reference and source views. Specifically, we take the first image as the reference view and predict all Gaussian in its camera frame. We use a reference embedding e^{ref} for the first view ($n = 1$) and another source embedding e^{src} for other views ($n = 2, \dots, N$), both of which are learnable.

Feed-forward Transformer. The augmented tokens undergo processing through L self-attention blocks, each combining self-attention and MLP layers with pre-normalization and residual connections [21].

Gaussian Map Prediction. Each image token $e_{\text{out}}^{n,m}$ output by the last self-attention block is transformed into p^2 Gaussians with a linear layer, yielding vectors of dimension $p^2 \cdot q$ (q being the Gaussian parameter count). These predictions are reshaped into Gaussian patches $G^{n,m} \in \mathbb{R}^{p \times p \times q}$ and spatially concatenated to form N Gaussian maps $\{G^n \in \mathbb{R}^{H \times W \times q} \mid n = 1, \dots, N\}$.

Each map pixel represents a q -dimensional 3D Gaussian primitive. Unlike pose-dependent Gaussian LRMs [44, 58, 65] that use single depth values for Gaussian locations, our pose-free setting precludes depth-based unprojection. Instead, we directly predict Gaussian locations in the reference frame and enforce pixel alignment through a dedicated loss term to restrict Gaussians to lie on camera rays (de-

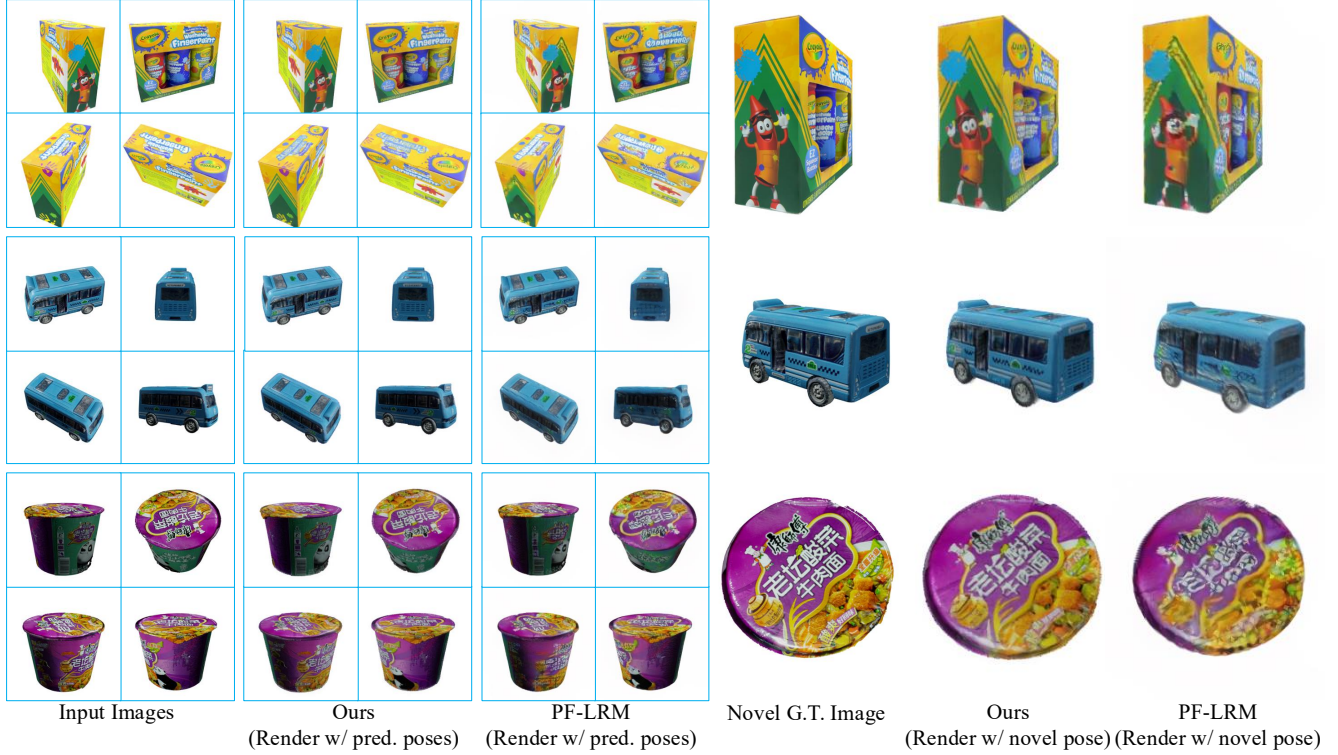


Figure 3. **Sparse-view Reconstruction on PF-LRM’s Evaluation Datasets.** FreeSplatter-O synthesizes significantly better visual details than PF-LRM. The 1st row is from the GSO dataset, while the 2nd and 3rd rows are from the OmniObject3D dataset.

tailed in Section 3.3).

Camera Pose Estimation. Camera parameter recovery begins with focal length f estimation from predicted Gaussian maps. Unlike DUS3R, which requires *pairwise* image processing and subsequent *global alignment*, FreeSplatter predicts all Gaussian maps in a unified reference frame, enabling direct camera pose estimation for all views. Given the n -th view’s Gaussian location map $\mathbf{X}^n \in \mathbb{R}^{H \times W \times 3}$ (first 3 channels of \mathbf{G}^n), corresponding pixel coordinate map $\mathbf{Y}^n \in \mathbb{R}^{H \times W \times 2}$, and validity mask $\mathbf{M}^n \in \mathbb{R}^{H \times W}$, we employ PnP-RANSAC [2, 20] to compute the camera pose $\mathbf{P}^n \in \mathbb{R}^{4 \times 4}$:

$$\mathbf{P}^n = \text{PnP}(\mathbf{X}^n, \mathbf{Y}^n, \mathbf{M}^n, \mathbf{K}), \quad (3)$$

where $\mathbf{K} = [[f, 0, \frac{W}{2}], [0, f, \frac{H}{2}], [0, 0, 1]]$ represents the estimated intrinsic matrix. The mask \mathbf{M}^n identifies valid pixels for pose optimization, which is implemented differently for object-centric and scene-level reconstruction. Please refer to Section 1 of the supplementary material for more implementation details.

3.3. Training Details

FreeSplatter offers two variants optimized for object-centric and scene-level pose-free reconstruction. While sharing architectural elements and parameter scale, these variants employ distinct training objectives and strategies.

Two-stage Training Strategy. Prior pose-dependent LRMs leverage pure rendering loss for supervision [23, 29, 58, 65]. However, our model assumes no known camera poses nor intrinsics and the Gaussian positions are free in 3D space, making it extremely challenging to predict correct Gaussian positions. Gaussian-based reconstruction approaches heavily rely on the initialization of Gaussian positions, *e.g.*, 3DGS [27] initializes the Gaussian positions with the sparse point cloud generated by SfM, while the parameters of our model are randomly initialized at the beginning. In practice, we found it essential to supervise the Gaussian positions at the beginning:

$$\mathcal{L}_{\text{pos}} = \sum_{n=1}^N \left\| \mathbf{M}^n \odot \hat{\mathbf{X}}^n - \mathbf{M}^n \odot \mathbf{X}^n \right\|, \quad (4)$$

where $\hat{\mathbf{X}}^n \in \mathbb{R}^{H \times W \times 3}$ represents predicted positions, \mathbf{X}^n denotes ground truth positions from depth unprojection, and $\mathbf{M}^n \in \mathbb{R}^{H \times W}$ masks valid depth values, which is the foreground object mask for object-centric reconstruction. For scene-level reconstruction, \mathbf{M}^n depends on where the depth values are defined in different datasets.

We apply \mathcal{L}_{pos} in the pre-training stage, so that the model learns to predict approximately correct Gaussian positions. In our experiments, this pre-training is *essential* to model’s convergence. However, \mathcal{L}_{pos} can only supervise the pixels with valid depths, while the Gaussian positions predicted



Figure 4. **Sparse-view Reconstruction on GSO dataset.** * indicates that ground truth camera poses are used as input.

at other pixels remain unconstrained. Besides, the ground truth depths are noisy in some datasets, and applying \mathcal{L}_{pos} throughout the training leads to degraded rendering quality. To provide a more stable geometric supervision, we adopt a pixel-alignment loss to enforce each predicted Gaussian to be aligned with its corresponding pixel through cosine similarity maximization:

$$\mathcal{L}_{\text{align}} = \sum_{n=1}^N \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left(1 - \frac{\hat{\mathbf{r}}_{i,j}^n \cdot \mathbf{r}_{i,j}^n}{\|\hat{\mathbf{r}}_{i,j}^n\| \|\mathbf{r}_{i,j}^n\|} \right), \quad (5)$$

where $\mathbf{r}_{i,j}^n$ denotes the ray from the camera origin \mathbf{t}^n to point $\mathbf{X}_{i,j}^n$. $\mathcal{L}_{\text{align}}$ restricts the predicted Gaussians to be distributed on the camera rays, which enhances rendering quality and facilitates camera parameter estimation by minimizing pixel-projection errors.

Loss Functions. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{render}} + \lambda_a \cdot \mathcal{L}_{\text{align}} + \mathbf{1}_{t \leq T_{\text{max}}} \lambda_p \cdot \mathcal{L}_{\text{pos}}, \quad (6)$$

where the rendering loss $\mathcal{L}_{\text{render}}$ is a combination of MSE and LPIPS loss. t and T_{max} denote the current training step and maximum pre-training step, respectively. In our implementation, we set $\lambda_a = 1.0$, $\lambda_p = 10.0$, $T_{\text{max}} = 10^5$.

Occlusion in Pixel-aligned Gaussians. Pose-dependent Gaussian-based LRMs [44, 58, 65] parameterize Gaussian positions with single depth values to ensure pixel alignment. Despite the simplicity, this approach limits reconstruction to areas directly observed in input views, potentially missing occluded regions in sparse-view scenarios. Our model addresses this limitation differently for object-centric and scene-level reconstruction: (i) For object-centric reconstruction, we apply $\mathcal{L}_{\text{align}}$ (Equation 5) exclusively to foreground regions, allowing Gaussians outside these areas to position freely and model occluded regions. (ii) For

Method	GSO			OmniObject3D		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Evaluate renderings at G.T. novel-view poses						
PF-LRM	25.08	0.877	0.095	21.77	0.866	0.097
FreeSplatter-O	23.54	0.864	0.100	22.83	0.876	0.088
Evaluate renderings at predicted input poses						
PF-LRM	27.10	0.905	0.065	25.86	0.901	0.062
FreeSplatter-O	25.50	0.897	0.076	26.49	0.926	0.050

Table 1. **Sparse-view Reconstruction on PF-LRM’s Eval Data.**

Method	GSO			
	RRE \downarrow	RRA@15° \uparrow	RRA@30° \uparrow	TE \downarrow
PF-LRM	3.99	0.956	0.976	0.041
FreeSplatter-O	8.96	0.909	0.936	0.090
OmniObject3D				
RRE \downarrow RRA@15° \uparrow RRA@30° \uparrow TE \downarrow				
PF-LRM	8.013	0.889	0.954	0.089
FreeSplatter-O	3.446	0.982	0.996	0.039

Table 2. **Camera Pose Estimation on PF-LRM’s Eval Data.**

scene-level reconstruction with real-world imagery, complete pixel alignment is necessary to handle complex backgrounds. We focus on reconstructing observed areas and adopt Splatt3R’s [42] target-view masking strategy, computing rendering loss only for visible regions to prevent negative training guidance from occluded areas.

4. Experiments

We evaluate our method on both sparse-view reconstruction (Section 4.2) and camera pose estimation (Section 4.3) tasks, including object-centric and scene-level scenarios. Please refer to the supplementary material for additional implementation details and experimental results.

4.1. Experimental Settings

Training Datasets. FreeSplatter-O is trained on Objaverse [9], utilizing white-background renders of centered objects. Each 3D asset is normalized to a $[-1, 1]^3$ cube, with 32 randomly sampled views (with diverse camera intrinsics, *i.e.*, focal lengths) and corresponding depth maps rendered at 512×512 resolution. FreeSplatter-S leverages a diverse training set comprising Blended-MVS [61], ScanNet++[62], and CO3Dv2[37]—a subset of DUST3R’s [51] training data encompassing outdoor scenes, indoor environments, and real-world objects.

Evaluation Datasets. For object-level experiments, we utilize Google Scanned Objects (GSO)[13] and OmniObject3D [56] (chosen 300 objects across 30 categories). Each object is captured through 24 views: 20 random and 4 structured input views, the latter positioned uniformly at 20° elevation for comprehensive coverage. In addition, we also use the GSO/OmniObject3D evaluation data provided by PF-LRM for comparison, since we can only access its in-

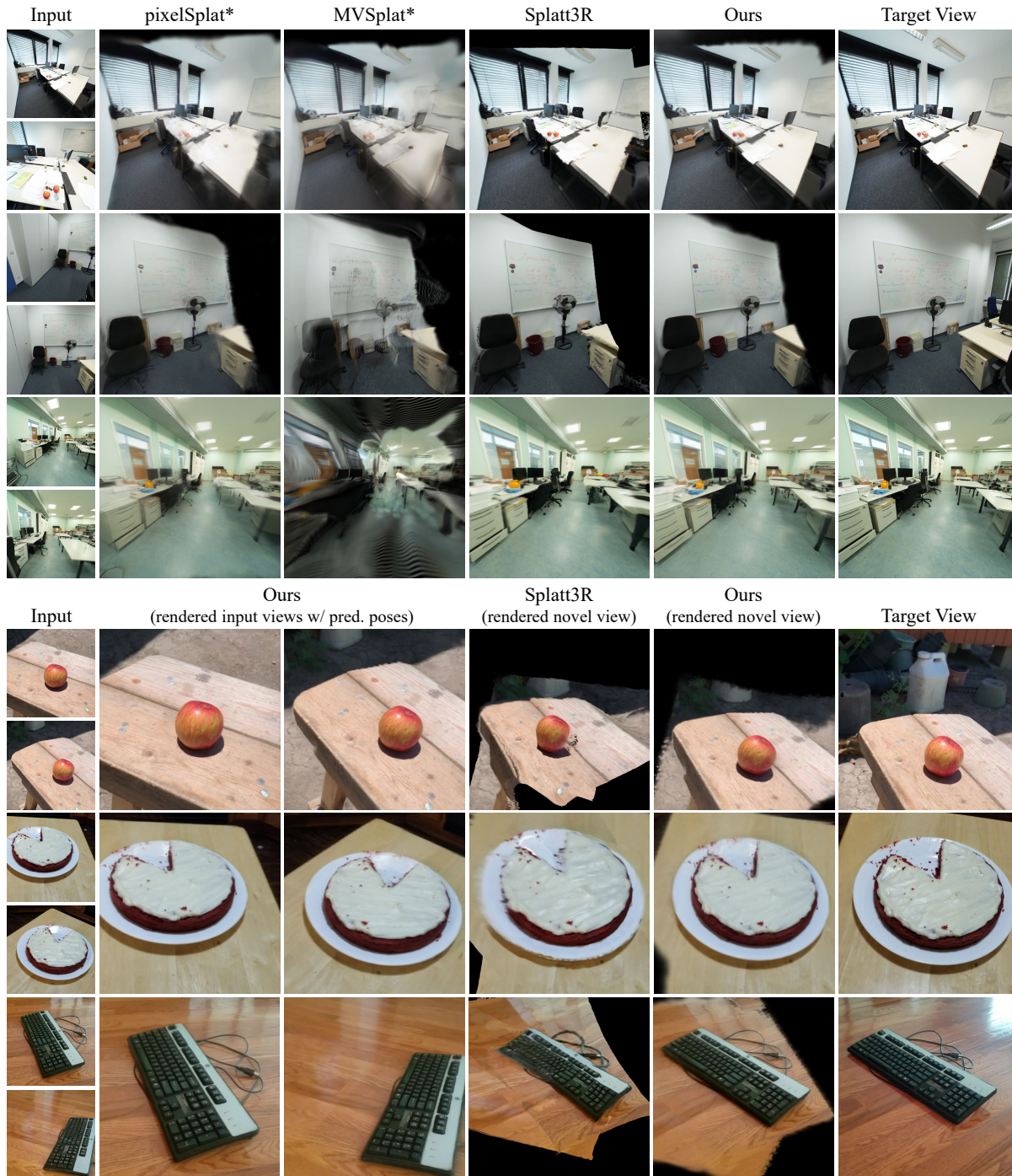


Figure 5. **Sparse-view Reconstruction on ScanNet++ (top) and CO3Dv2 (bottom).** * indicates that ground truth camera poses are used as input.

ference results. Scene-level performance is assessed on the test splits of ScanNet++[62] and CO3Dv2 [37].

4.2. Sparse-view Reconstruction

Baselines. Prior pose-free object reconstruction approaches like LEAP [26] exhibits limited generalization due to its small-scale training, while PF-LRM [49] is highly rele-

vant and serves as our baseline for both object-level reconstruction and pose estimation tasks. We also evaluate against two pose-dependent methods LGM [44] and InstantMesh [57], which leverage 3D Gaussians and tri-plane NeRF respectively, using ground truth camera poses. For scene-level reconstruction, we compare against two state-of-the-art generalizable Gaussian methods: pixelSplat [5]

Method	GSO			OmniObject3D		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
LGM*	24.463	0.891	0.093	24.852	0.942	0.060
InstantMesh*	25.421	0.891	0.095	24.077	0.945	0.062
FreeSplatter-O	30.443	0.945	0.055	31.929	0.973	0.027

Method	ScanNet++			CO3Dv2		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
pixelSplat*	24.974	0.889	0.180	-	-	-
MVSplat*	22.601	0.862	0.208	-	-	-
Splatt3R	21.013	0.830	0.209	18.074	0.740	0.197
FreeSplatter-S	25.807	0.887	0.140	20.405	0.781	0.162

Table 3. **Sparse-view Reconstruction on Object-centric and Scene-level Datasets.** We did not test pixelSplat/MVSplat on CO3Dv2 due to the significant domain gap. * indicates that ground truth camera poses are used as input.

and MVSplat [8]. Both methods are fine-tuned on ScanNet++ after pre-training on RealEstate10K [67]. We also evaluate against Splatt3R [42], a pose-free approach that combines a frozen MAST3R [28] backbone with a trainable head for Gaussian attribute prediction.

Metrics. We evaluate the performance of sparse-view reconstruction using standard novel view synthesis metrics (PSNR, SSIM, and LPIPS) at 512 \times 512 resolution.

Comparison with PF-LRM. Due to the lack of code, we benchmark against PF-LRM using their provided evaluation datasets and inference results. As Table 1 shows, while PF-LRM achieves superior metrics on their GSO evaluation dataset, FreeSplatter-O performs better on their OmniObject3D evaluation dataset. This disparity can be attributed to PF-LRM’s GSO evaluation images being rendered under identical conditions (*e.g.*, light intensity, camera distribution) as their training data, whereas OmniObject3D uses original dataset images, providing a more objective comparison. Qualitative results in Figure 3 demonstrate FreeSplatter’s superior preservation of visual details.

Comparison with Pose-dependent LRMs. On our object-centric evaluation datasets, FreeSplatter-O significantly outperforms pose-dependent methods LGM and InstantMesh, achieving PSNR improvements of > 5 and > 7 on GSO and OmniObject3D respectively, despite their usage of ground truth camera poses (Table 3). Qualitative comparisons in Figure 4 reveal superior detail preservation by our method, particularly evident in text rendering (4th column), while competitors exhibit blurring artifacts. Existing works on LRMs assume the necessity of accurate camera poses for high-quality 3D reconstruction, incorporating pose information through LayerNorm modulation [29] or plucker ray embeddings [44, 58, 59]. However, FreeSplatter-O’s superior performance suggests that scalable and high-quality sparse-view reconstruction is feasible without known accurate camera poses in certain cases.

Results on Scene-level Reconstruction. For scene-level reconstruction, FreeSplatter-S outperforms pose-dependent

Method (Object)	GSO			
	RRE \downarrow	RRA@15 $^\circ$ \uparrow	RRA@30 $^\circ$ \uparrow	TE \downarrow
FORGE	97.814	0.022	0.083	0.898
MASt3R	59.633	0.269	0.456	0.353
FreeSplatter-O	3.902	0.961	0.977	0.040

Method (Object)	OmniObject3D			
	RRE \downarrow	RRA@15 $^\circ$ \uparrow	RRA@30 $^\circ$ \uparrow	TE \downarrow
FORGE	76.822	0.081	0.257	0.430
MASt3R	91.204	0.105	0.212	0.524
FreeSplatter-O	11.346	0.909	0.935	0.104

Method (Scene)	ScanNet++			
	RRE \downarrow	RRA@15 $^\circ$ \uparrow	RRA@30 $^\circ$ \uparrow	TE \downarrow
RoMa	0.862	0.977	0.985	0.421
MASt3R	0.724	0.988	0.993	0.356
FreeSplatter-S	0.776	0.991	0.990	0.066

Method (Object)	CO3Dv2			
	RRE \downarrow	RRA@15 $^\circ$ \uparrow	RRA@30 $^\circ$ \uparrow	TE \downarrow
PoseDiffusion	7.950	0.803	0.868	0.409
RayDiffusion	7.028	0.833	0.890	0.482
RoMa	5.377	0.839	0.922	0.335
MASt3R	2.917	0.975	0.989	0.299
FreeSplatter-S	3.048	0.976	0.986	0.190

Method (Object)	Re10K			
	RRE \downarrow	RRA@15 $^\circ$ \uparrow	RRA@30 $^\circ$ \uparrow	TE \downarrow
PoseDiffusion	14.387	0.732	0.780	0.466
RayDiffusion	12.023	0.767	0.814	0.439
RoMa	5.663	0.918	0.947	0.402
MASt3R	2.341	0.972	0.994	0.374
FreeSplatter-S	3.513	0.982	0.995	0.293

Table 4. **Camera Pose Estimation on Object-centric and Scene-level Datasets.** To be noted, Re10K is outside the training dataset.

methods (pixelSplat, MVSplat) on most ScanNet++ metrics (Table 3). While Splatt3R, a pose-free alternative, employs MAST3R’s [28] frozen architecture for point prediction, its performance is limited by fixed Gaussian positions. Our end-to-end training approach enables joint optimization of Gaussian parameters, resulting in superior visual fidelity on both ScanNet++ and CO3Dv2 datasets (Figure 5). To be noted, novel view synthesis for pose-free methods is accomplished through camera alignment with target viewpoints.

4.3. Camera Pose Estimation

Baselines. We first evaluate pose estimation performance against PF-LRM on its evaluation datasets. For all of our evaluation datasets, we benchmark against MAST3R, the current state-of-the-art in zero-shot multi-view pose estimation. Additional comparisons include FORGE [25] for object-centric evaluation, and PoseDiffusion [48], RayDiffusion [64], and RoMa [16] for scene-level tasks, with the former two excluded from ScanNet++ evaluation due to training scope limitations. Traditional COLMAP-based methods [40] are omitted due to documented high failure rates in sparse-view scenarios [49]. We further incorporate RealEstate10K [67] (Re10K) test splits to assess generalization to challenging scenes.

Metrics. Following established protocols [48, 49], we eval-

# Layers (L)	Patch Size (P)	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
16	16	25.417	0.896	0.088
16	8	28.945	0.934	0.064
24	16	28.622	0.927	0.063
24	8	30.443	0.945	0.055

Table 5. **Ablation Study on Model Architecture.** The results are evaluated on the GSO dataset with FreeSplatter-O.

uate pose estimation performance using both rotation and translation metrics: relative rotation error (RRE) in degrees, relative rotation accuracy (RRA) at 15° and 30° thresholds, and translation error (TE) measured as the distance between predicted and ground truth camera centers. For multi-view settings, errors are averaged over all possible pairs of cameras. It is important to note that the TE metric is scale-invariant: we first compute the relative translations between views for both ground truth and predictions, *normalize* these translations by their respective mean ℓ_2 -norm, and then report the mean difference.

Comparison with PF-LRM. Pose estimation results mirror reconstruction performance trends: PF-LRM excels on their GSO evaluation set, while FreeSplatter-O demonstrates superior performance on OmniObject3D. As we have analyzed, this disparity likely stems from PF-LRM’s GSO evaluation images sharing characteristics with their training data, making OmniObject3D a more objective benchmark.

Comparison on Our Evaluation Datasets. Table 4 demonstrates the significant performance advantage of FreeSplatter-O over existing baselines on object-centric datasets. MAST3R’s reduced effectiveness in this context can be attributed to domain gaps between its training data and background-free rendered images. In scene-level evaluation, FreeSplatter-S matches or exceeds MAST3R’s performance, showing superior RRA@ 15° and TE metrics on ScanNet++ and CO3Dv2. Notably, FreeSplatter-S achieves state-of-the-art performance on the challenging Re10K benchmark despite utilizing a smaller training corpus compared to MAST3R.

4.4. Ablation Studies

Model Architecture. We analyze architectural choices using a base configuration of 24 transformer layers with patch size 8. Table 5 demonstrates consistent performance improvements on GSO with increased layer count and reduced patch size, attributed to enhanced model capacity and reduced information loss, respectively.

View Embedding Addition. We evaluate the impact of view embedding addition as formulated in Equation 2. Experiments reveal that assigning e^{ref} to the j -th view’s tokens and e^{src} to remaining views’ tokens enables successful reference view identification and accurate Gaussian reconstruction in the corresponding camera frame. Alternative embedding combinations result in degraded reconstruction quality (details in Section 2.7 of supplementary material).

$\mathcal{L}_{\text{align}}$	GSO			ScanNet++		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
\times	26.684	0.898	0.092	21.330	0.832	0.201
\checkmark	30.443	0.945	0.055	25.807	0.887	0.140

Table 6. **Ablation Study on Pixel-alignment Loss.** The results on GSO and ScanNet++ are evaluated with FreeSplatter-O and FreeSplatter-S, respectively.

Number of Input Views. We conduct an experiment on the GSO dataset to illustrate how the number of input views influences the reconstruction quality. Please refer to Figure 13 of the supplementary material for more details.

Pixel-Alignment Loss. Ablation on the pixel-alignment loss (Equation 5) demonstrates its crucial role in both object and scene-level reconstruction. Its removal leads to significant degradation across all metrics on GSO and ScanNet++ datasets (Table 6). Figure 12 in the supplementary material illustrates how this loss term enhances visual fidelity, with its absence resulting in notable blur artifacts.

4.5. Applications in 3D AIGC

FreeSplatter integrates seamlessly into 3D content creation pipelines, offering substantial operational advantages through its pose-free architecture. In contrast, traditional pipelines [29, 44, 52, 57] require precise alignment between the camera configurations of multi-view diffusion models and the parameters of LRMs, which introduces complexity and potential sources of error. FreeSplatter removes these constraints, enabling direct processing of multi-view images without the need for camera pose information. This streamlined workflow not only reduces generation time for users but also maintains—or even improves—reconstruction quality. In our supplementary material (Section 2.4), we provide comprehensive image-to-3D generation results across a range of multi-view diffusion models, demonstrating that FreeSplatter achieves superior reconstruction performance compared to pose-dependent LRMs and can accurately recover predefined camera parameters from diffusion model outputs.

5. Conclusion

FreeSplatter presents a scalable framework for pose-free sparse-view reconstruction. Leveraging a single-stream transformer architecture and unified-frame Gaussian map prediction, the framework delivers both high-fidelity 3D reconstruction and efficient camera pose estimation. Its two specialized variants, designed for object-centric and scene-level reconstruction, achieve superior performance in terms of both reconstruction quality and pose accuracy. Additionally, FreeSplatter shows significant potential in boosting the productivity of downstream applications such as text/image-to-3D content creation, freeing users from the complexities associated with camera pose handling.

References

- [1] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint arXiv:2408.00653*, 2024. 2
- [2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4
- [3] Ang Cao, Justin Johnson, Andrea Vedaldi, and David Novotny. Lightplane: Highly-scalable components for neural 3d fields. *arXiv preprint arXiv:2404.19760*, 2024. 2
- [4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2
- [5] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 1, 2, 6
- [6] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. *arXiv preprint arXiv:2407.04699*, 2024. 2
- [7] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2781–2790, 2022. 2
- [8] 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. *arXiv preprint arXiv:2403.14627*, 2024. 2, 7
- [9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 5
- [10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3
- [13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 5
- [14] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 2
- [15] Johan Edstedt, Ioannis Athanasiadis, Márten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 2
- [16] Johan Edstedt, Qiyu Sun, Georg Bökman, Márten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 2, 7
- [17] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Hanwen Jiang, Dejia Xu, Zehao Zhu, Dilin Wang, and Zhangyang Wang. Pose-free generalizable rendering transformer. *arXiv e-prints*, pages arXiv–2310, 2023. 2
- [18] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024. 2
- [19] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2
- [20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 4
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [22] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jiaolong Yang, Seungryong Kim, and Chong Luo. Unifying correspondence pose and nerf for generalized pose-free novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20196–20206, 2024. 2
- [23] 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 *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 4
- [24] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically ac-

- curate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
- [25] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *2024 International Conference on 3D Vision (3DV)*, pages 31–41. IEEE, 2024. 2, 7
- [26] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. LEAP: Liberate sparse-view 3d modeling from camera poses. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6
- [27] 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. 1, 2, 3, 4
- [28] Vincent Leroy, Johann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2, 7
- [29] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4, 7, 8
- [30] Mengfei Li, Xiaoxiao Long, Yixun Liang, Weiyu Li, Yuan Liu, Peng Li, Xiaowei Chi, Xingqun Qi, Wei Xue, Wenhan Luo, et al. M-Irm: Multi-view large reconstruction model. *arXiv preprint arXiv:2406.07648*, 2024. 2
- [31] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5741–5751, 2021. 2
- [32] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 2
- [33] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, et al. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024. 2
- [34] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Fast generalizable gaussian splatting reconstruction from multi-view stereo. *arXiv preprint arXiv:2405.12218*, 2024. 2
- [35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2
- [36] Frank Plautia. The weiszfeld algorithm: proof, amendments, and extensions. *Foundations of location analysis*, pages 357–389, 2011. 2
- [37] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 2, 5, 6
- [38] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 2
- [39] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2
- [40] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2, 7
- [41] Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *arXiv preprint arXiv:2407.02445*, 2024. 2
- [42] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 2, 5, 7
- [43] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10208–10217, 2024. 2
- [44] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 2, 3, 5, 6, 7, 8
- [45] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 2
- [46] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2
- [47] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Visual geometry grounded deep structure from motion. *arXiv preprint arXiv:2312.04563*, 2023. 2
- [48] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 7
- [49] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. PF-LRM: Pose-free large reconstruction model for joint pose and shape prediction. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6, 7
- [50] Shuzhe Wang, Juho Kannala, Marc Pollefeys, and Daniel Barath. Guiding local feature matching with surface curvature. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17981–17991, 2023. 2

- [51] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. [2](#), [5](#)
- [52] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. [2](#), [8](#)
- [53] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepssf: Structure from motion via deep bundle adjustment. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 230–247. Springer, 2020. [2](#)
- [54] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. [2](#)
- [55] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024. [2](#)
- [56] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. [5](#)
- [57] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. [1](#), [2](#), [6](#), [8](#)
- [58] 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. *arXiv preprint arXiv:2403.14621*, 2024. [2](#), [3](#), [4](#), [5](#), [7](#)
- [59] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. DMV3d: Denoising multi-view diffusion using 3d large reconstruction model. In *The Twelfth International Conference on Learning Representations*, 2024. [2](#), [7](#)
- [60] Fan Yang, Jianfeng Zhang, Yichun Shi, Bowen Chen, Chenxu Zhang, Huichao Zhang, Xiaofeng Yang, Jiashi Feng, and Guosheng Lin. Magic-boost: Boost 3d generation with mutli-view conditioned diffusion. *arXiv preprint arXiv:2404.06429*, 2024. [2](#)
- [61] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. [2](#), [5](#)
- [62] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. [2](#), [5](#), [6](#)
- [63] Chubin Zhang, Hongliang Song, Yi Wei, Yu Chen, Jiwen Lu, and Yansong Tang. Geolrm: Geometry-aware large reconstruction model for high-quality 3d gaussian generation. *arXiv preprint arXiv:2406.15333*, 2024. [2](#)
- [64] Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. [7](#)
- [65] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024. [2](#), [3](#), [4](#), [5](#)
- [66] Xin-Yang Zheng, Hao Pan, Yu-Xiao Guo, Xin Tong, and Yang Liu. Mvd²: Efficient multiview 3d reconstruction for multiview diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. [2](#)
- [67] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph*, 37, 2018. [7](#)