

RegGS: Unposed Sparse Views Gaussian Splatting with 3DGS Registration

Chong Cheng^{1*} Yu Hu^{1*} Sicheng Yu¹ Beizhen Zhao¹ Zijian Wang¹ Hao Wang^{1†}

¹The Hong Kong University of Science and Technology (Guangzhou)

{ccheng735, yhu847}@connect.hkust-gz.edu.cn haowang@hkust-gz.edu.cn

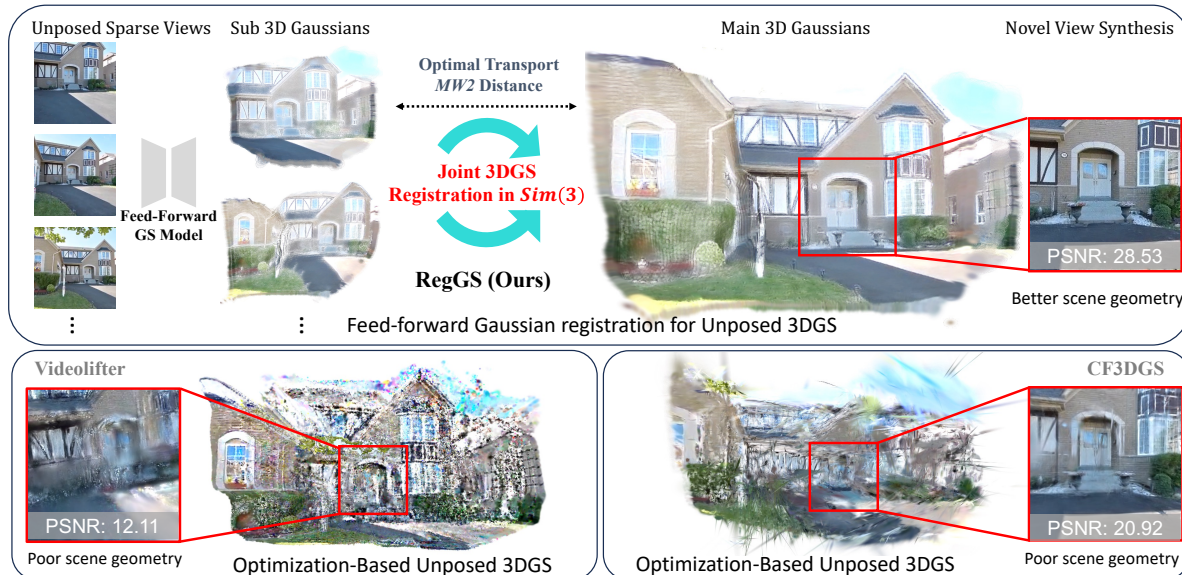


Figure 1. Overview of our pipeline for 3D Gaussian Splatting from multiple unposed sparse views. A pre-trained feed-forward GS model extracts sub 3D Gaussians from each input, while two initial images yield the main 3D Gaussians. We measure the structural closeness of Gaussian sets using the entropy-regularized MW_2 distance and align them in $Sim(3)$ space with our joint 3DGS registration module. Our method outperforms others in reconstruction quality and novel view synthesis.

Abstract

3D Gaussian Splatting (3DGS) has demonstrated its potential in reconstructing scenes from unposed images. However, optimization-based 3DGS methods struggle with sparse views due to limited prior knowledge. Meanwhile, feed-forward Gaussian approaches are constrained by input formats, making it challenging to incorporate more input views. To address these challenges, we propose RegGS, a 3D Gaussian registration-based framework for reconstructing unposed sparse views. RegGS aligns local 3D Gaussians generated by a feed-forward network into a globally consistent 3D Gaussian representation. Technically, we implement an entropy-regularized Sinkhorn algorithm to efficiently solve the optimal transport Mixture 2-Wasserstein (MW_2) distance, which serves as an alignment metric for Gaussian mixture models (GMMs) in $Sim(3)$ space. Fur-

thermore, we design a joint 3DGS registration module that integrates the MW_2 distance, photometric consistency, and depth geometry. This enables a coarse-to-fine registration process while accurately estimating camera poses and aligning the scene. Experiments on the RE10K and ACID datasets demonstrate that RegGS effectively registers local Gaussians with high fidelity, achieving precise pose estimation and high-quality novel-view synthesis. Project page: <https://3dagentworld.github.io/reggs/>.

1. Introduction

Recent advances in 3D reconstruction and novel view synthesis—driven by the demand for immersive experiences in VR, AR, and robotics—have yielded impressive results under dense observations [7, 15, 17, 34, 38]. Reconstructing 3D scenes from sparse, unposed data remains a formidable challenge, as real-world conditions often provide limited overlap and unreliable camera poses [10].

* Equal contribution.

† Corresponding author.

Despite the effectiveness of Neural Radiance Fields (NeRF) [27] in novel view synthesis, traditional NeRF methods often require known camera poses [3, 25, 28, 35, 42], limiting their broader application. Recent efforts to combine pose estimation with NeRF [4, 8, 23, 36] face issues of difficult convergence and high computational costs. Optimization-based 3D Gaussian Splatting (3DGS) [6, 18, 21, 26, 29] methods have shown potential in real-time scene reconstruction but struggle with sparse views due to insufficient geometric priors. These limitations often lead to topological discontinuities and scale ambiguities, significantly reducing their practicality.

In contrast, feedforward-based methods [5, 9, 19, 39, 40, 44] leverage implicit 3D priors learned from large-scale training data, enabling direct prediction of coherent 3D Gaussians from images without iterative optimization. This learned prior not only enhances cross-dataset generalization but also regularizes the reconstruction in scenarios with under-constrained geometric information [11, 41]. Recent approaches [32, 40] achieve direct inference of 3D Gaussian representations from unposed images, eliminating the need for iterative optimization.

However, feed-forward methods can only handle a limited number of input images, restricting their applicability to broader scenarios. This raises an intriguing question: *Can we register locally generated Gaussian models from a feed-forward network into a globally consistent 3D Gaussian representation?*

To address this issue, we propose a novel 3D Gaussian reconstruction framework: **RegGS**, which performs unposed sparse view reconstruction by registering feed-forward Gaussian incrementally. Specifically, we introduce the optimal transport-based Mixture 2-Wasserstein (MW_2) distance between Gaussian mixture models (GMM) to align generalized Gaussian manifolds. Through a differentiable multi-modal joint registration pipeline, we solve for scene alignment in the $\text{Sim}(3)$ space.

Technically, we utilize the entropy-regularized Sinkhorn algorithm to compute the differentiable upper bound MW_2 for the W_2 distance between GMMs, thereby circumventing the infinite-dimensional W_2 optimization problem. By integrating engineering techniques such as log-Sinkhorn and Cholesky decomposition, we efficiently compute the MW_2 distance between thousands of 3D Gaussians on GPU, thereby accurately measuring their alignment in the $\text{Sim}(3)$ space.

Furthermore, we incorporate the global distribution of the MW_2 distance, photometric consistency, and depth geometry into a joint 3D Gaussian registration module, enabling elastic scale alignment and topology adaptation within $\text{Sim}(3)$. By performing a coarse-to-fine incremental 3DGS registration followed by global optimization, we achieve high-precision camera pose estimation and high-

quality scene reconstruction. Our contribution can be summarized as:

- We construct an optimal transport framework for Gaussian Mixture Models in the $\text{Sim}(3)$ space and efficiently compute the MW_2 distance using the entropy-regularized Sinkhorn algorithm, thereby providing a differentiable alignment metric for 3D Gaussian distributions.
- We propose a 3DGS joint registration module that achieves precise camera pose estimation and scene registration by jointly utilizing MW_2 distance, photometric consistency, and depth geometry.
- Experiments on the RE10K and ACID datasets demonstrate that RegGS significantly improves pose estimation accuracy and the quality of novel view synthesis, offering broad possibilities for practical applications.

2. Related Work

2.1. NeRF-based Pose-Free Reconstruction

Novel view synthesis, particularly in the absence of accurate camera poses, has garnered significant attention in recent years. Traditional Neural Radiance Fields (NeRF) methods [3, 25, 27, 42] have achieved remarkable results. However, these methods usually rely on known camera poses for training, limiting their applicability in scenarios where pose information is unavailable or unreliable which is very common in real-world scenarios.

Several approaches have been proposed to extend NeRF to handle unposed input images. Among them, [8, 12] integrate camera pose estimation with NeRF rendering, leveraging a recurrent GRU module for pose and depth estimation. Similarly, [33] employs a weighted Procrustes analysis and an optical flow network to establish correspondences for pose estimation. More recently, CoPoNeRF [20] introduced a unified framework that integrates correspondence matching, pose estimation, and NeRF rendering, allowing for end-to-end training and improved performance in challenging scenarios with extreme viewpoint changes. Additionally, methods like Nope-NeRF [4] leverage depth information to constrain the optimization process.

While NeRF-based methods show promise, their reliance on dense ray sampling leads to slow training and inference, struggles with extreme viewpoint changes and minimal overlap, and high computational costs.

2.2. Optimization-based Pose-Free 3DGS Reconstruction

3D Gaussian Splatting (3DGS) [21] offers an alternative by representing the scene with a set of 3D Gaussians, which can be rendered efficiently. However, traditional 3DGS also relies on accurate camera poses and sparse point clouds from Structure-from-Motion (SfM) pipelines like Colmap.

To address this, Colmap-Free 3DGS [18] proposes a

method to optimize the 3D Gaussian representation directly from unposed images. By incorporating pose estimation into the optimization loop, this approach eliminates the need for precomputed poses, making it more flexible and applicable to a wider range of scenarios.

Similarly, videoLifter uses pre-trained models [22, 37] to reconstruct globally consistent 3D models from uncalibrated monocular videos, reducing error accumulation and computational costs. Yet, it struggles with sparse view reconstruction challenges. While optimization-based methods can achieve high-quality reconstructions, they struggle to efficiently handle sparse viewpoint scenes and face challenges in learning complex 3D spatial relationships.

2.3. Feedforward-based Pose-Free 3DGS Reconstruction

Feed-forward approaches aim to alleviate this by predicting the 3D representation directly from the input images in a single pass. NoPoSplat [40] exemplifies this by using a neural network to map unposed images to a 3D Gaussian representation in a canonical space, enabling fast and efficient reconstruction without iterative optimization. Other feed-forward methods, such as pixelSplat [5] and MVSplat [9], predict Gaussian primitives from posed images, leveraging geometric priors like epipolar geometry or cost volumes.

In contrast, NoPoSplat operates without poses by directly predicting Gaussians in a canonical space, demonstrating improved performance, especially in scenarios with limited overlap between input views. However, feed-forward Gaussian models typically handle only a limited number of input images, limiting their application in scenarios with large coverage and sparse viewpoints.

Consequently, we explored a method based on 3D Gaussian registration to achieve incremental unposed sparse view reconstruction. This approach not only leverages the excellent scene priors of feed-forward models but also enables high-quality reconstruction in broader sparse view scenarios, which is of practical importance.

3. Method

As shown in Fig. 2, our method initializes a main map from two images using a pretrained feed-forward Gaussian model, and generates sub Gaussians for each subsequent image. By measuring similarity between the GMMs through an optimal transport MW_2 distance by an entropy-regularized Sinkhorn approach, our differentiable joint 3DGS registration module estimates the Sim(3) transformation before merging local Gaussians into the main map. Finally, we perform a global refinement of the 3D Gaussians with adaptive pruning, yielding high-fidelity reconstructions even from unposed sparse views.

3.1. Registration Problem Modeling

The core of our work is 3DGS registration. An intuitive approach is to use 3DGS center points as registration references. However, these center points cannot accurately reflect the geometric structure of the scene. Here, we introduce a statistical model, Gaussian Mixture Model (GMM) [16], which can describe the structural distribution of 3D Gaussians based on their attributes. Specifically, we first define the main 3D Gaussians between two frames, with the main Gaussians \mathcal{G}^A and sub Gaussians \mathcal{G}^B expressed as GMMs:

$$\mathcal{G}^A = \sum_{i=1}^M w_i^A \mathcal{N}(\mu_i^A, \Sigma_i^A), \quad (1)$$

$$\mathcal{G}^B = \sum_{k=1}^N w_k^B \mathcal{N}(\mu_k^B, \Sigma_k^B), \quad (2)$$

where μ represents the mean of the Gaussian distribution, Σ represents the covariance matrix, and weights satisfy $\sum_i w_i^A = 1$, $\sum_k w_k^B = 1$, obtained through opacity normalization.

It is notable that we do not consider color information (spherical harmonic coefficients), as color information is unstable due to lighting angle variations. Our goal is to find the optimal **affine transformation** $T \in \text{Sim}(3)$ parameters, including rotation $R \in SO(3)$, translation $t \in \mathbb{R}^3$, and scaling factor $s \in \mathbb{R}^+$, such that the structural difference between the transformed sub Gaussians $T(\mathcal{G}^B)$ and the main Gaussians \mathcal{G}^A is minimized. The objective function is:

$$T^* = \arg \min_{T \in \text{Sim}(3)} \mathcal{D}(\mathcal{G}^A, T(\mathcal{G}^B)), \quad (3)$$

where \mathcal{D} is a distance metric function used to measure the difference between two sets of 3D Gaussian distributions. After the Sim(3) transformation of the sub-map, the parameters of each Gaussian component change according to the following relationships:

$$\mu_k^{B'} = s R \mu_k^B + t, \quad \Sigma_k^{B'} = s^2 R \Sigma_k^B R^\top. \quad (4)$$

Under the above transformation, we compute the matching relationship between Gaussian components in the main Gaussians \mathcal{G}^A and the transformed sub Gaussians $T(\mathcal{G}^B)$ by minimizing the \mathcal{D} distance.

3.2. Optimal Transport MW_2 Distance

Inspired by previous research [1], we adopt the 2-Wasserstein (W_2) distance as the fundamental metric to measure geometric differences between two sets of 3D Gaussian distributions. For two Gaussian components $\mathcal{N}(\mu_i^A, \Sigma_i^A)$ and $\mathcal{N}(\mu_k^{B'}, \Sigma_k^{B'})$, the square of their W_2 distance is defined as:

$$W_2^2 = |\mu_i^A - \mu_k^{B'}|^2 + \text{Tr} \left(\Sigma_i^A + \Sigma_k^{B'} - 2 \left(\Sigma_i^A \Sigma_k^{B'} \right)^{1/2} \right), \quad (5)$$

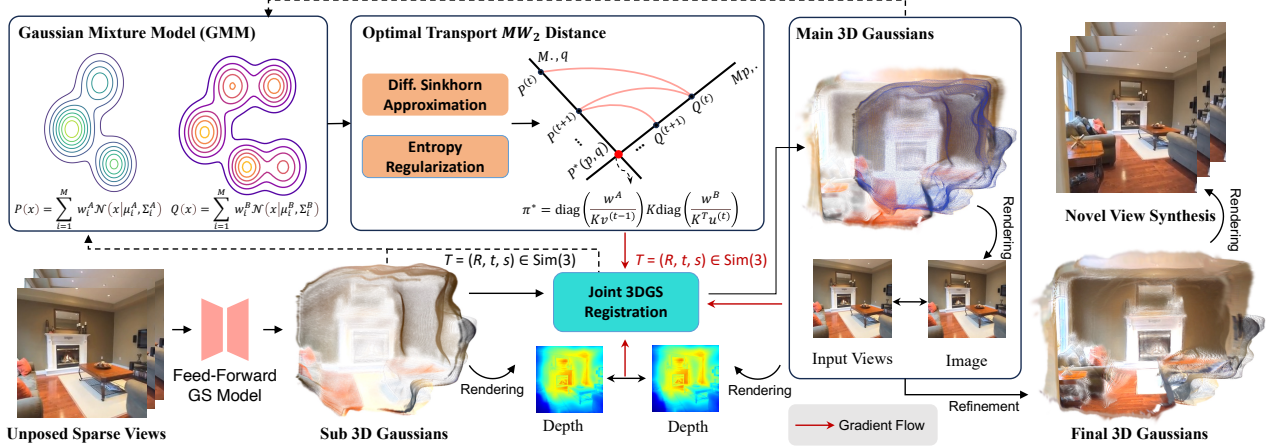


Figure 2. **Pipeline of Unposed Sparse Views Gaussian Splatting with 3DGS Registration (RegGS).** First, we use a pre-trained feed-forward Gaussian model to construct a main Gaussians from two initial images. Then, for each new input, a sub Gaussians is generated and aligned with the main Gaussians. Specifically, by solving the optimal transport MW_2 distance with an entropy-regularized Sinkhorn approximation, our differentiable 3DGS joint registration module estimates the $\text{Sim}(3)$ transformation and merges the sub Gaussians into the main Gaussians. Finally, we perform refinement of the global Gaussians, yielding a high-fidelity 3D reconstruction.

where the position term $\|\mu_i^A - \mu_k^{B'}\|^2$ reflects the Euclidean offset between distribution centers, and the covariance term eliminates rotation effects through matrix square roots, becoming zero when $\Sigma_i^A = \Sigma_k^{B'}$.

However, directly computing the W_2 distance between GMMs requires solving an infinite-dimensional optimization problem, which is computationally infeasible [2]. To address this, we introduce the ‘‘GMM transport’’ method, which constrains the optimal transport plan to the Gaussian mixture subspace, transforming the continuous problem into a discrete linear assignment problem [16]. Its mathematical form is:

$$MW_2^2(P, Q) = \inf_{\pi \in \Pi(w^A, w^B)} \sum_{i=1}^M \sum_{k=1}^N \pi_{ik} C_{ik}, \quad (6)$$

where C_{ik} is the transport cost for the Gaussian pair (i, k) , and $\Pi(w^A, w^B)$ is the set of transport plans satisfying $\sum_i \pi_{ik} = w_k^B$ and $\sum_k \pi_{ik} = w_i^A$. In this case, MW_2 forms an upper bound of W_2 , satisfying $MW_2(\mu_0, \mu_1) \geq W_2(\mu_0, \mu_1)$ [16].

We employ the optimal transport Sinkhorn algorithm [14] to compute the MW_2 distance. Since the two sets of Gaussian spheres are not in one-to-one correspondence and are numerous, to avoid local minima, accelerate convergence, and enable fuzzy matching, we employ an entropy regularization strategy to construct a differentiable Sinkhorn approximation. The optimization objective is:

$$W_{2,\epsilon}^2 = \min_{\pi \in \Pi(w^A, w^B)} \left[\sum_{i,k} \pi_{ik} C_{ik} + \epsilon \sum_{i,k} \pi_{ik} \log \pi_{ik} \right], \quad (7)$$

where ϵ controls the regularization strength. We solve this problem through Sinkhorn iterations: initially, we initialize the kernel matrix $K_{ik} = \exp(-C_{ik}/\epsilon)$; subsequently, we alternately perform scaling updates:

$$u^{(t)} = \frac{w^A}{K v^{(t-1)}}, \quad v^{(t)} = \frac{w^B}{K^\top u^{(t)}}. \quad (8)$$

After T iterations, we obtain the transport plan

$$\pi^* = \text{diag}(u^{(T)}) K \text{diag}(v^{(T)}), \quad (9)$$

and finally calculate the entropy-regularized Wasserstein distance

$$W_{2,\epsilon}^2 = \sum_{i,k} \pi_{ik}^* C_{ik}. \quad (10)$$

This method reduces the computational complexity to $O(MN)$ while ensuring gradient differentiability. The proof of gradient consistency for the entropy-regularized Sinkhorn W_2 distance, along with the complexity calculations, can be found in the appendix.

3.3. Differentiable Joint 3DGS Registration

To establish an efficient and stable 3D Gaussian registration model, we propose a differentiable framework based on quaternion parameterization and multi-objective joint optimization. In traditional methods, pose parameterization often faces redundancy or singularity issues, and our proposed Sinkhorn approximation of MW_2 distance is not an exact solution, making single-objective optimization prone to local optima. Therefore, we design a strategy that integrates quaternion pose representation, multi-loss joint optimization, and adaptive weight allocation, with mathematical formulation and implementation details as follows.

Pose Parameterization Design: We represent a Sim(3) transformation by decomposing it into a quaternion rotation $\mathbf{q} \in S^3$, a translation $\mathbf{t} \in \mathbb{R}^3$, and a logarithmic scale $\log s \in \mathbb{R}$, forming the parameter vector $\boldsymbol{\theta} = [\mathbf{q}; \mathbf{t}; \log s] \in \mathbb{R}^8$. This formulation guarantees positive scaling via $s = \exp(\log s)$ and enforces $\|\mathbf{q}\| = 1$ using projected gradient updates. When applied to Gaussian components, the update formulas for mean and covariance are:

$$\begin{aligned} \mu_k^{B'} &= s \cdot R(\mathbf{q})\mu_k^B + \mathbf{t}, \\ \Sigma_k^{B'} &= s^2 \cdot R(\mathbf{q})\Sigma_k^B R(\mathbf{q})^\top, \end{aligned} \quad (11)$$

where the rotation matrix $R(\mathbf{q})$ is analytically generated from the quaternion $\mathbf{q} = [w, x, y, z]^\top$:

$$R(\mathbf{q}) = \begin{bmatrix} 1 - 2y^2 - 2z^2 & 2xy - 2wz & 2xz + 2wy \\ 2xy + 2wz & 1 - 2x^2 - 2z^2 & 2yz - 2wx \\ 2xz - 2wy & 2yz + 2wx & 1 - 2x^2 - 2y^2 \end{bmatrix}. \quad (12)$$

Our experiments show that quaternion rotation converges significantly faster than Lie algebra rotation while achieving equivalent accuracy.

Multi-Loss Joint Optimization: To balance global distribution alignment and precise geometric consistency, we construct a joint loss function:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{MW}_2} + \lambda_2 \mathcal{L}_{\text{Photo}} + \lambda_3 \mathcal{L}_{\text{Depth}}, \quad (13)$$

where the global alignment term $\mathcal{L}_{\text{MW}_2} = W_{2,\epsilon}^2(G^A, T(G^B))$ is calculated using the differentiable Sinkhorn algorithm from Sec. 3.2, driving the overall matching of Gaussian distribution centers and covariances; the local photometric term uses the 3DGS differentiable rendering pipeline [21] to generate RGB images from aligned viewpoints, enhancing precise map alignment through pixel-level L1 loss. The local photometric loss is described as:

$$\mathcal{L}_{\text{Photo}} = \frac{1}{|P|} \sum_{p \in P} |I^A(p) - I^{T(B)}(p)|_1, \quad (14)$$

where $T(G^B)$ represents applying the current Sim(3) transformation with parameters $\boldsymbol{\theta}$ to the source distribution G^B ; depth is similarly rendered using the 3DGS differentiable rendering pipeline [21], with invalid regions excluded through an effective depth mask M_v , suppressing scale drift and topological distortion. The depth geometric constraint term is described as:

$$\mathcal{L}_{\text{Depth}} = \frac{1}{|M_v|} \sum_{p \in M_v} |D_A^v(p) - D_{T(B)}^v(p)|, \quad (15)$$

where $D_A^v(p) \in \mathbb{R}^+$ and $D_{T(B)}^v(p) \in \mathbb{R}^+$ are depth maps under viewpoint v , and M_v is the valid depth mask.

Differentiable Gradient Path: To achieve end-to-end optimization, we calculate the gradient of the loss with respect

to parameters $\boldsymbol{\theta}$. For the MW_2 term, its gradient propagates through the transport plan π_{ik}^* and the chain rule:

$$\frac{\partial \mathcal{L}_{\text{MW}_2}}{\partial \boldsymbol{\theta}} = \sum_{i,k} \pi_{ik}^* \left(\frac{\partial C_{ik}}{\partial \mu_k^{B'}} \frac{\partial \mu_k^{B'}}{\partial \boldsymbol{\theta}} + \frac{\partial C_{ik}}{\partial \Sigma_k^{B'}} \frac{\partial \Sigma_k^{B'}}{\partial \boldsymbol{\theta}} \right), \quad (16)$$

where the Jacobian matrix of quaternion rotation $\partial R(\mathbf{q})/\partial \mathbf{q}$ is implicitly solved by automatic differentiation. The gradients of photometric and depth terms are back-propagated through the rendering pipeline:

$$\frac{\partial \mathcal{L}_{\text{Photo}}}{\partial \boldsymbol{\theta}} = \frac{1}{|P|} \sum_p \text{sign}(I^A - I^{T(B)}) \cdot \frac{\partial I^{T(B)}}{\partial \mu_k^{B'}} \frac{\partial \mu_k^{B'}}{\partial \boldsymbol{\theta}}, \quad (17)$$

$$\frac{\partial \mathcal{L}_{\text{Depth}}}{\partial \boldsymbol{\theta}} = \frac{1}{|M_v|} \sum_{p \in M_v} \text{sign}(D_A^v - D_{T(B)}^v) \cdot \frac{\partial D_{T(B)}^v}{\partial \mu_k^{B'}} \frac{\partial \mu_k^{B'}}{\partial \boldsymbol{\theta}}, \quad (18)$$

where the rendering gradients $\partial I/\partial \mu_k^{B'}$ and $\partial D/\partial \mu_k^{B'}$ are analytically derived from the 3DGS volume rendering formula [21].

The joint optimization of these three components allows for fast and robust registration of 3DGS sub-maps. Subsequently, the next frame is inferred as a sub-map by the pre-trained model, continuously updating the main map to complete the reconstruction.

3.4. Joint Training

Joint 3DGS Registration. Feed-forward Gaussian models often produce targets with vastly different scales. To avoid falling into local optima, we perform scale normalization before optimization. We begin by calculating the average value of depth rendered from sub Gaussian map, which is generated by the feed-forward Gaussian model, denoted as D_{sub} , and scale it to a common scale. Moreover, in joint optimization, to enhance the efficiency of iterative optimization, initialization is also necessary. We compare the depth values of the main Gaussians function D_{main} with those of the sub Gaussians function D_{sub} to determine the initial relative scale s_{init} .

Computational Efficiency. To achieve efficient computation of large-scale Gaussian MW_2 distances, we map Sinkhorn iteration operations, including matrix scaling, covariance matrix Cholesky decomposition, and Wasserstein distance calculation to GPU through tensorized operations, achieving efficient computation between Gaussian pairs through batch parallel processing. To address the risk of exponential term overflow in entropy regularization, we design a logarithmic space accumulation strategy that maintains numerical stability when computing MW_2 , while uniformly regularizing covariance matrices as $\Sigma \leftarrow \Sigma + 10^{-6}I$ to ensure positive definiteness.

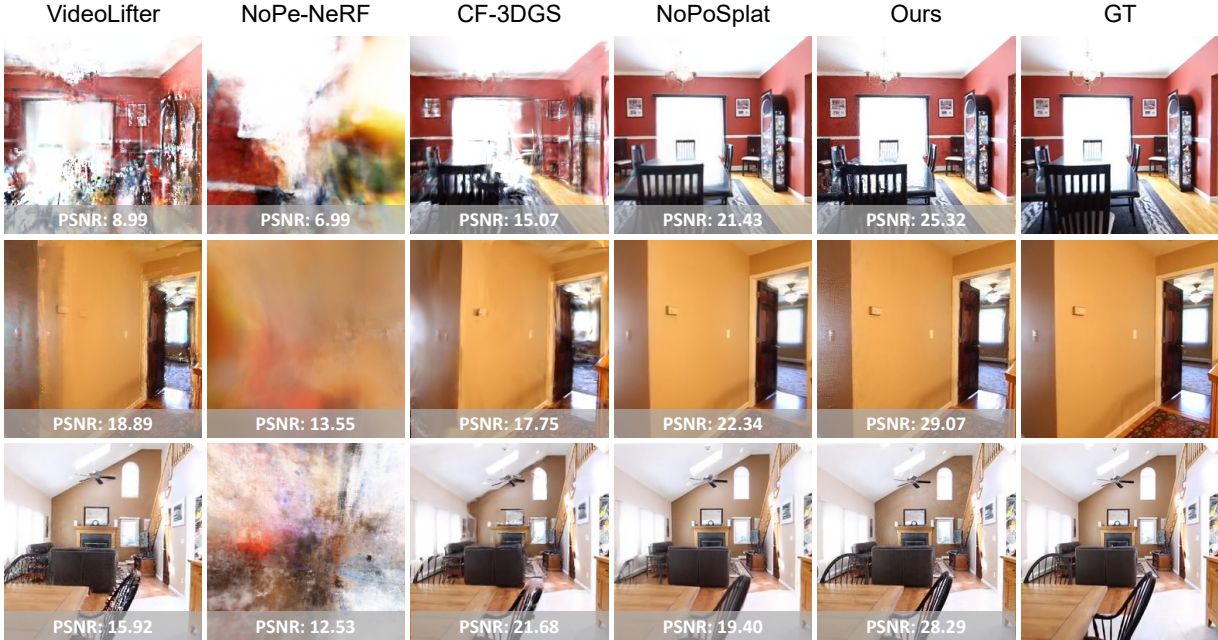


Figure 3. **Qualitative Comparison on the RE10K [43]**. NoPoSplat: 2× views; others: 16× views. Our method not only registers the 3D Gaussians but also enhances novel view synthesis through global refinement.

Method	2×			8×			16×			32×		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
COLMAP* [30]	9.687	0.266	0.533	7.171	0.135	0.676	18.904	0.614	0.294	22.911	0.725	0.219
Splatt3R [32]	13.951	0.442	0.443	-	-	-	-	-	-	-	-	-
NoPoSplat [40]	23.247	0.832	0.111	-	-	-	-	-	-	-	-	-
CF-3DGS [18]	19.326	0.638	0.277	20.329	0.672	0.235	23.034	0.792	0.188	25.596	0.865	0.133
NoPeNeRF [4]	10.225	0.351	0.781	10.974	0.343	0.767	10.465	0.321	0.763	10.021	0.284	0.742
VideoLifter [13]	14.526	0.448	0.346	16.651	0.564	0.273	14.765	0.452	0.382	15.268	0.483	0.344
MASt3R* [22]	16.036	0.580	0.361	24.249	0.824	0.189	27.024	0.869	0.149	28.309	0.891	0.094
RegGS (Ours)	24.272	0.853	0.174	26.691	0.877	0.185	28.663	0.913	0.147	28.332	0.912	0.151

Table 1. **Novel View Synthesis Results on the RE10K [43]**. The terms “2x”, “8x”, “16x”, and “32x” represent the number of views in the input images. An asterisk (*) indicates reconstruction with 3DGS. A dash (-) indicates that the input is not supported by the method. Our method outperforms other unposed methods in reconstruction quality with sparse views, and the gap widens as the number of views decreases.

4. Experiment

4.1. Experiment Setup

Datasets. To evaluate the effectiveness of our method, we conducted experiments on the RE10K [43] and ACID [24] datasets. The RE10K dataset includes indoor and outdoor scene videos, while ACID consists mainly of aerial shots of natural landscapes captured by drones. Both provide camera poses and intrinsic parameters. Following the setup in [40], we use the test sets of each dataset for evaluation.

For the unposed sparse views reconstruction task, the number of views we reconstructed are 2, 8, 16, and 32. To simulate sparse input, both training and testing views are equidistantly sampled from the videos. For 2-view scenar-

ios, we sample every 40 frames for videos with significant motion and every 60 frames for scenes with less motion. For scenarios with 8, 16, and 32 views, training views are equidistantly sampled throughout the entire video. The test set includes all frames not used for training.

Evaluation Metrics. To evaluate novel view synthesis (NVS), we use PSNR, SSIM, and LPIPS as metrics. For pose estimation evaluation, we use ATE RMSE as a metric. For 3DGS registration evaluation, we use the MW_2 distance. As illustrated in Fig. 5, the proposed MW_2 distance precisely quantifies the proximity between two GMMs.

Baselines. We compare our method with methods for unposed reconstruction in the NVS task, including: Colmap [30, 31], NoPoSplat [40], NoPe-NeRF [4], VideoLifter

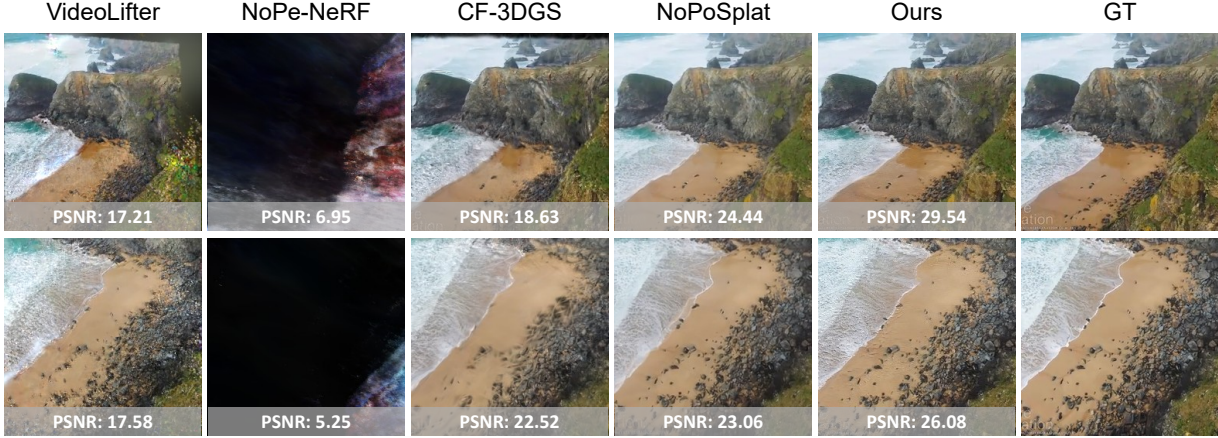


Figure 4. **Qualitative comparison on the ACID [24]**. NoPoSplat: 2x views; others: 16x views. Our method is applicable to both indoor scenes and drone-captured videos, demonstrating superior novel view synthesis performance.

Method	2x			8x			16x			32x		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
COLMAP* [30]	8.340	0.141	0.643	14.162	0.207	0.554	7.904	0.049	0.719	7.300	0.058	0.716
Splatt3R [32]	10.468	0.215	0.591	-	-	-	-	-	-	-	-	-
NoPoSplat [40]	23.589	0.663	0.202	-	-	-	-	-	-	-	-	-
CF-3DGS [18]	21.654	0.604	0.301	22.212	0.629	0.289	23.458	0.651	0.266	23.419	0.650	0.263
NoPeNerf [4]	13.231	0.269	0.748	14.611	0.273	0.732	6.837	0.117	0.788	11.961	0.222	0.756
VideoLifter [13]	17.921	0.327	0.405	18.830	0.332	0.394	18.264	0.289	0.412	19.503	0.393	0.335
MASt3R* [22]	18.390	0.312	0.447	22.231	0.525	0.318	24.537	0.673	0.240	25.216	0.702	0.155
RegGS (Ours)	24.291	0.703	0.237	25.764	0.753	0.252	27.745	0.834	0.201	26.772	0.774	0.243

Table 2. **Novel View Synthesis Results on the ACID [24]**. The terms “2x”, “8x”, “16x”, and “32x” represent the number of views in the input images. An asterisk (*) indicates reconstruction with 3DGS. A dash (-) indicates that the input is not supported by the method. The data shows that our method also outperforms other unposed reconstruction methods in drone-captured scenes. As the scene becomes sparser, the gap between our method and the others increases.

Method	RE10K			ACID		
	8x ATE \downarrow	16x ATE \downarrow	32x ATE \downarrow	8x ATE \downarrow	16x ATE \downarrow	32x ATE \downarrow
VideoLifter	0.335	0.291	0.232	0.272	0.206	0.145
NoPeNerf	0.844	0.902	0.597	0.684	0.413	0.455
CF3DGS	0.237	0.254	0.286	0.278	0.195	0.239
Ours	0.023	0.041	0.078	0.020	0.038	0.095

Table 3. **Pose estimation results on the RE10K [43] and ACID [24]**. We evaluate the pose estimation accuracy of our method with different numbers of input views. Our method outperforms other baseline methods in terms of pose accuracy.

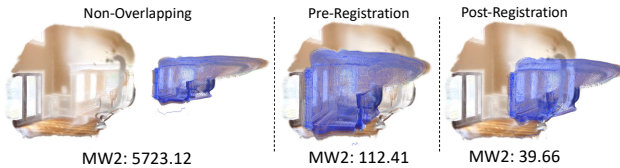


Figure 5. MW_2 distances effectively quantify alignment levels between sets of 3D Gaussians under various conditions. Notably, the rightmost case aligns with the correct position.

[13], CF-3DGS [18], MARSt3R [22], and Splatt3R [32].

Implementation Details. The hardware used in our experiments is the NVIDIA A6000. Our method is implemented using PyTorch, with NoPoSplat [40] as the backbone. In the pose estimation of training frames and the scale estimation of sub Gaussians, we perform joint optimization. After completing registration and optimization for all frames, we perform global refinement to further refine the scene.

4.2. Experimental Results and Analysis

Novel View Synthesis: As shown in Tab. 1, Tab. 2, Fig. 3 and Fig. 4, our method significantly outperforms other unposed reconstruction methods in terms of PSNR and SSIM. NoPe-NeRF [4] fails to converge; VideoLifter [13] produces distorted renderings under sparse views; and CF-3DGS [18] suffers from artifacts due to inadequate detail capture. For LPIPS, we generally lead, though we occasionally fall short in some cases, due to noise introduced by global refinement when improving PSNR.

Pose Estimation: Our method can also be applied to pose

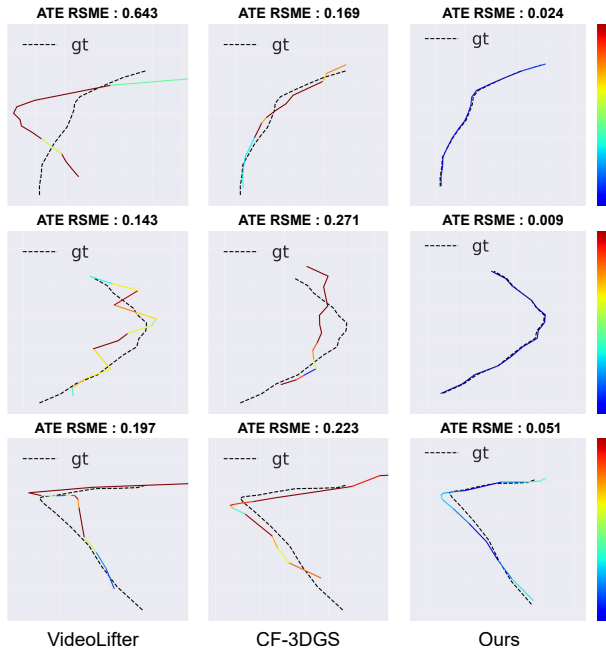


Figure 6. **Trajectory Comparison on the RE10K [43].** Our method and the baseline are under 16-view input. Our method achieves higher pose estimation accuracy than other unposed methods and is applicable to various scenes and camera motions.

estimation. As shown in Tab. 3 and Fig. 6. We conduct experiments on the RE10K [43] and ACID [24] datasets under 8, 16, and 32-view input conditions. The poses estimated by the baseline methods were aligned with the ground truth (GT) poses for comparison. Table 3 presents the performance of our method. Compared to other unposed methods, our method demonstrates a more pronounced performance gap, especially in sparse view conditions.

4.3. Ablation Studies

Ablation Study on Loss Function: In this section, we investigate the 3DGS joint optimization loss function described in Sec. 3.3. To validate the performance of our designed loss function, we conduct experiments on the RE10K [43] dataset by testing the results when each individual loss term is omitted. The input is set to 16 views, and the evaluation metrics used for comparison are ATE, PSNR, SSIM, LPIPS, and MW_2 . To facilitate the comparison of the MW_2 loss, we normalize its values to a range of 0 to 100, representing the baseline for convergence.

As shown in Fig. 5, the MW_2 distance measures the closeness of the Gaussian scene structure distribution. Experiments in Tab. 4 demonstrate that the MW_2 loss supports coarse alignment and pose estimation but may lead to local minima and misalignment when used alone. Photometric loss is essential for refining registration and improving NVS, yet it may cause submaps to converge to separate spatial regions. Depth-consistency loss stabilizes pose and ge-

	ATE↓	PSNR↑	SSIM↑	LPIPS↓	MW_2 ↓
w/o Photo	1.184	16.06	0.52	0.44	58.8
w/o Depth	0.160	20.97	0.72	0.29	57.8
w/o MW_2	1.151	19.41	0.67	0.31	67.7
RegGS (Ours)	0.098	23.09	0.79	0.23	56.5

Table 4. **Ablations on Loss Functions.** The performance of our method degrades when any loss term is removed, demonstrating the effectiveness of the loss functions we employ.

	ATE↓	PSNR↑	SSIM↑	LPIPS↓	MW_2 ↓
w/o JR	1.164	11.41	0.34	0.60	100.0
RegGS (Ours)	0.098	23.09	0.79	0.23	56.5

Table 5. **Ablations on key Modules.** The results show that precise pose estimation and 3DGS registration depend on the 3DGS joint registration (JR) module.

ometry but fails to converge in isolation. These results underscore the necessity of jointly optimizing all loss terms for accurate registration.

Ablation Study on Key Module: The key module in our approach is the joint 3DGS registration. We perform experiments following the same setup as in the previous experiments. As shown in Tab. 5, when the 3DGS joint registration module is removed, there is a significant decline in scene reconstruction and pose estimation accuracy, indicating the critical role of this module in accurate pose estimation and 3DGS registration.

4.4. Limitations

Our method is influenced by the performance of feed-forward Gaussians; poor quality generation by these models can lead to registration and fusion failures. Additionally, the training time increases significantly with more input views due to the MW_2 distance, indicating the need for further optimization. In cases of large inter-frame motion, the registration process may also fail to converge.

5. Conclusion

This paper presents RegGS, an incremental 3D Gaussian reconstruction framework for unposed sparse view settings. We constructed a GMM alignment metric in $\text{Sim}(3)$ space based on the optimal transport MW_2 distance, and efficiently computed the MW_2 distance using the entropy-regularized Sinkhorn algorithm, thereby circumventing the infinite-dimensional optimization problem. By jointly optimizing MW_2 , photometric, and depth-consistency losses, RegGS achieves progressive coarse-to-fine registration of both camera poses and scene structure. Experiments on RE10K and ACID demonstrate superior pose estimation and novel view synthesis compared to prior methods, highlighting RegGS’s potential for real-world applications.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (No. 62406267), Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3956 & Grant No.2023A03J0008), the Guangzhou Municipal Science and Technology Project (No. 2025A04J4070), and the Guangzhou Municipal Education Project (No. 2024312122).

References

- [1] Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters can be computed in polynomial time in fixed dimension. *Journal of Machine Learning Research*, 22(44): 1–19, 2021. 3
- [2] Jason M. Altschuler and Enric Boix-Adserà. Wasserstein barycenters are np-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022. 4
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2
- [4] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023. 2, 6, 7
- [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. 2, 3
- [6] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2
- [7] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024. 1
- [8] Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24–34, 2023. 2
- [9] 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 *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 2, 3
- [10] Chong Cheng, Gaochao Song, Yiyang Yao, Qinzhen Zhou, Gangjian Zhang, and Hao Wang. Graph-guided scene reconstruction from images with 3d gaussian splatting, 2025. 1
- [11] Chong Cheng, Sicheng Yu, Zijian Wang, Yifan Zhou, and Hao Wang. Outdoor monocular slam with global scale-consistent 3d gaussian pointmaps, 2025. 2
- [12] Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhansu Maji, and Ameesh Makadia. Lu-nerf: Scene and pose estimation by synchronizing local unposed nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18312–18321, 2023. 2
- [13] Wenyan Cong, Kevin Wang, Jiahui Lei, Colton Stearns, Yuanhao Cai, Dilin Wang, Rakesh Ranjan, Matt Feiszli, Leonidas Guibas, Zhangyang Wang, Weiyao Wang, and Zhiwen Fan. Videolifter: Lifting videos to 3d with fast hierarchical stereo alignment, 2025. 6, 7
- [14] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc. 4
- [15] Anurag Dalal, Daniel Hagen, Kjell G Robbersmyr, and Kristian Muri Knausgård. Gaussian splatting: 3d reconstruction and novel view synthesis, a review. *IEEE Access*, 2024. 1
- [16] Julie Delon and Agnès Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020. 3, 4
- [17] Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1
- [18] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiao-long Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20796–20805, 2024. 2, 6, 7
- [19] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. Pf3plat: Pose-free feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2410.22128*, 2024. 2
- [20] 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
- [21] 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. 2, 5
- [22] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 3, 6, 7
- [23] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2
- [24] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6, 7, 8
- [25] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances*

- in *Neural Information Processing Systems*, 33:15651–15663, 2020. 2
- [26] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 2
- [27] 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. 2
- [28] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. 2
- [29] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *arXiv preprint arXiv:2403.17898*, 2024. 2
- [30] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6, 7
- [31] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 6
- [32] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs, 2024. 2, 6, 7
- [33] Cameron Smith, Yilun Du, Ayush Tewari, and Vincent Sitzmann. Flowcam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow. *arXiv preprint arXiv:2306.00180*, 2023. 2
- [34] Gaochao Song, Chong Cheng, and Hao Wang. Gvkf: Gaussian voxel kernel functions for highly efficient surface reconstruction in open scenes, 2024. 1
- [35] Matthew Tancik, Vincent Casser, Xinchun Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8248–8258, 2022. 2
- [36] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses, 2023. 2
- [37] 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 (CVPR)*, pages 20697–20709, 2024. 3
- [38] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4):613–642, 2024. 1
- [39] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthspat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 2
- [40] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 2, 3, 6, 7
- [41] Sicheng Yu, Chong Cheng, Yifan Zhou, Xiaojun Yang, and Hao Wang. Rgb-only gaussian splatting slam for unbounded outdoor scenes, 2025. 2
- [42] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2
- [43] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 6, 7, 8
- [44] 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, 2024. 2