SalientGS: Unified SfM-to-3DGS with Importance-Guided MCMC Gaussian Allocation

Tianyu Xiong
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

Rui Li
CEMSE, King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Suning Ge
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

Jiaqi Yang
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

Abstract

Reconstructing 3D scenes from unordered images remains bottlenecked by expensive Structure-from-Motion (SfM) preprocessing and frozen pose interfaces. We present SalientGS, a unified SfM-to-3D Gaussian Splatting (3DGS) pipeline. Its central contribution is importance-guided Markov Chain Monte Carlo (MCMC) Gaussian allocation, which aggregates multi-view residuals into per-Gaussian underfit and redundancy signals. These signals define a smooth importance-weighted sampling distribution that biases both birth and relocation toward underfit regions. This reallocates capacity from well-fit areas without altering the underlying stochastic gradient Langevin dynamics (SGLD). In a released-code verification over 13 scenes and three benchmarks, SalientGS achieves the best cross-benchmark macro-average PSNR, SSIM, and LPIPS (27.65 dB / 0.876 / 0.147) among the compared methods, while also providing the fastest end-to-end runtime (10.62 minutes) with 1.5M Gaussians. Code, per-scene measurements, and evaluation scripts are available at https://github.com/Six-Bit-TX/SalientGS.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

3D Gaussian Splatting, Structure from Motion, Joint Pose Optimization, Markov Chain Monte Carlo, Importance-Guided Allocation

ACM Reference Format:
Tianyu Xiong, Rui Li, Suning Ge, and Jiaqi Yang. 2026. SalientGS: Unified SfM-to-3DGS with Importance-Guided MCMC Gaussian Allocation. In *Proceedings of ACM Multimedia 2026 (ACM MM '26)*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM MM '26, Rio de Janeiro, Brazil

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

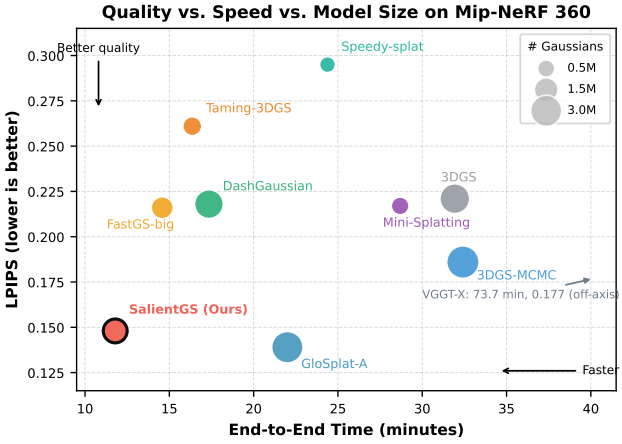

Figure 1: Quality vs. Speed vs. Model Size on Mip-NeRF 360. Bubble size indicates Gaussian count. SalientGS combines strong perceptual quality with the fastest end-to-end runtime in this comparison, using 1.5M Gaussians and no standalone COLMAP preprocessing.

1 Introduction

Reconstructing high-fidelity 3D scenes from unordered image collections is a fundamental problem in computer vision and graphics, with applications spanning virtual reality, robotics, and cultural heritage preservation. The recent advent of 3D Gaussian Splatting (3DGS) [16] has transformed this landscape, enabling real-time novel view synthesis with quality rivaling Neural Radiance Fields (NeRF) [24] while dramatically reducing rendering time. However, 3DGS inherits a critical dependency from the NeRF paradigm: it requires accurate camera poses and sparse point clouds from Structure-from-Motion (SfM) preprocessing, most commonly performed with COLMAP [30].

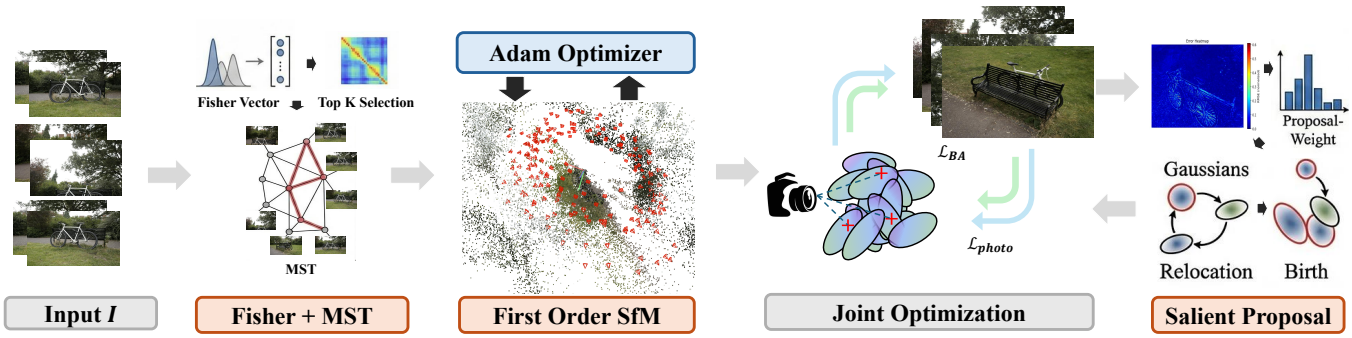

Figure 2: Starting from an unordered image set, we first retrieve candidate pairs with Fisher Vector descriptors and enforce global connectivity with a maximum spanning tree to build a sparse but reliable matching graph. We then run a first-order SfM stage to estimate initial camera poses and triangulate a coarse 3D structure, followed by joint optimization of the 3D Gaussian representation and camera parameters under photometric and reprojection-based BA losses. This unified pipeline concentrates model capacity on salient scene regions while preserving geometric anchoring throughout optimization. Red marks indicate the main novel components of our method.

In widely used COLMAP-based two-stage pipelines, this preprocessing step can become a practical bottleneck for end-to-end reconstruction. Exhaustive image matching and second-order bundle adjustment can be expensive, especially for large image collections, and their cost can rival or exceed 3DGS training time itself (Table 1). More importantly, the interface is typically frozen: SfM pose errors propagate downstream without photometric correction, and the separation between stages limits joint optimization of geometry and appearance.

Recent efforts to accelerate 3DGS training [5, 6, 22, 29] have achieved impressive speedups through improved densification strategies, progressive training, and efficient Gaussian management. However, these methods still rely on standalone COLMAP preprocessing, meaning their reported times understate the true end-to-end cost. Meanwhile, COLMAP-free approaches [7, 12, 14] have emerged, but many assume sequential input, use feed-forward networks that can be sensitive to domain shift, or omit geometric constraints during training, leading to pose drift in challenging scenarios.

We present **SalientGS**, a unified SfM-to-3DGS pipeline that combines fast reconstruction with competitive rendering quality under a fixed 1.5M-Gaussian budget. Our primary algorithmic contribution is *importance-guided MCMC Gaussian allocation*. We compute multi-view underfit (importance) and well-fit (redundancy) signals, then convert them into an importance-weighted sampling distribution. This distribution biases **both** birth and relocation within the MCMC framework, reallocating a fixed Gaussian budget from redundant regions to persistent errors. Crucially, this importance guidance operates as a heuristic allocation strategy layered on top of the SGLD-based population dynamics of 3DGS-MCMC [17]; it does not alter the underlying Langevin updates and therefore makes no additional convergence claims beyond those of the base framework. Our key insight is that *importance-guided capacity reallocation* makes a fast but coarse SfM initialization viable when it is coupled with joint refinement of pose and appearance and with geometric anchoring.

Unlike prior joint optimization methods that rely primarily on photometric gradients, SalientGS maintains explicit geometric constraints through a reprojection-based bundle adjustment (BA) loss on triangulated feature tracks, enabling accurate pose refinement while preventing degenerate solutions. Our contributions:

- **Importance-guided MCMC Gaussian allocation:**
  A heuristic allocation layer atop 3DGS-MCMC [17] that biases both birth and relocation toward underfit regions via multi-view error attribution, yielding +0.10 dB PSNR and a 0.001 LPIPS reduction over vanilla MCMC at 1.5M Gaussians (Table 2).
- **Unified SfM-to-3DGS pipeline:** An end-to-end architecture coupling fast SfM initialization with joint refinement of pose and appearance under photometric and reprojection losses.
- **Efficient matching and first-order SfM:** Fisher Vector (FV) retrieval with Maximum Spanning Tree (MST) connectivity and first-order epipolar adjustment, achieving near-linear scaling and up to 23× SfM speedup.
- **Released-code verification:** A reproducible 13-scene evaluation with exact 30K-step schedules and per-scene records; SalientGS gives the best three-benchmark macro-average PSNR/SSIM/LPIPS and runtime among all methods in Table 1.

2 Related Work

2.1 Novel View Synthesis and 3D Gaussian Splatting

NeRF [24] represents scenes as continuous volumetric functions, with improvements in anti-aliasing [3, 4] and training speed [26]. 3DGS [16] enables real-time rendering via anisotropic Gaussians but is sensitive to initialization quality. Recent acceleration efforts include importance-based budgeting [22], progressive reconstruction [5], pruning [6, 9], and MCMC-based management [17] with SGLD-driven birth-death processes. FastGS [29] combines many of these innovations. We adopt the MCMC framework and layer *importance-guided* birth and relocation on top, biasing allocation via multi-view error attribution without modifying the underlying SGLD dynamics.

2.2 Structure from Motion

Incremental SfM pipelines such as COLMAP [30] are widely used and robust, but they can accumulate drift and often rely on computationally heavy matching and second-order BA. Global SfM methods [25, 34] solve all poses simultaneously via rotation [10, 40] and translation averaging [8, 23]; GLOMAP [27] achieves COLMAP-level accuracy with significant speedups. FastMap [20] replaces second-order BA with first-order structureless epipolar adjustment, achieving BA-quality refinement with cost independent of point count. Learning-based approaches include differentiable SfM [37], flow-based optimization [33], and the DUSt3R/MASt3R/VGGT paradigm [19, 36, 38]. We integrate first-order SfM with joint 3DGS training, retaining an optimization-based geometric backbone while avoiding feed-forward domain sensitivity.

2.3 Image Matching and Retrieval

Exhaustive pairwise matching scales as $O(N^2)$. Retrieval-based pair selection addresses this via image-level descriptors such as Fisher Vectors [28], Bag-of-Words [13, 32], and learned methods like NetVLAD [1] and MegaLoc [2]. We adopt Fisher Vector retrieval with MST-based connectivity guarantees to achieve near-linear matching complexity without requiring pretrained networks.

2.4 Joint Pose and Appearance Optimization

Traditional pipelines treat SfM and novel view synthesis as independent modules with frozen interfaces. NeRF– [39] and BARF [21] optimize poses via photometric gradients; SPARF [35] adds multi-view correspondences but lacks explicit geometric constraints. COLMAP-free 3DGS methods have also emerged: CF-3DGS [7] and HT-3DGS [14] assume sequential input, 3RGS [12] relies on photometric-only refinement, and GloSplat [41] preserves SfM feature tracks for geometric anchoring. Unlike these methods, we maintain geometric constraints through a reprojection-based BA loss on triangulated tracks within a unified end-to-end pipeline.

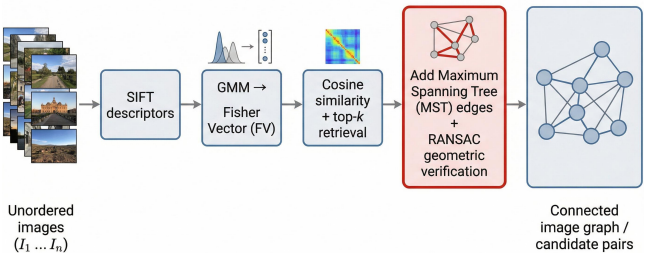

Figure 3: Given an unordered image collection, we extract Scale-Invariant Feature Transform (SIFT) features, encode them into Fisher Vectors, retrieve top-k candidate neighbors, add MST edges, and verify pairs with RANSAC. This produces a sparse but well-connected matching graph that balances reconstruction quality and efficiency.

3 Problem Description and Methodology

Given N unordered images depicting a scene, we simultaneously reconstruct a 3DGS representation and recover camera poses through three stages (Figure 2): global matching, first-order SfM, and joint 3DGS training with importance-guided MCMC allocation.

3.1 Global Correspondence with Fisher Vector & MST

Traditional exhaustive image matching scales as $O(N^2)$, becoming prohibitive for large image collections. We address this through a retrieval-based approach that identifies visually similar image pairs using Fisher Vector (FV) global descriptors, followed by Maximum Spanning Tree (MST) connectivity guarantees. Figure 3 illustrates our retrieval and pair selection pipeline.

Fisher Vector Encoding. For each image I_i, we extract Scale-Invariant Feature Transform (SIFT) descriptors and encode them into a Fisher Vector $\mathbf{f}_i$ [28] by computing gradients of the descriptor log-likelihood with respect to an M-component Gaussian Mixture Model (GMM) trained on the image collection. The resulting vectors are L2- and signed-square-root normalized for retrieval.

Pair Selection with MST Connectivity. We retrieve the top-k most similar images per query via FAISS [15] approximate nearest-neighbor (ANN) search in $O(N \log N)$ time, producing a sparse kNN graph. To guarantee connectivity for global SfM, we add Maximum Spanning Tree edges from this sparse graph, then apply RANSAC-based geometric verification to filter spurious matches. The final pair set is $\mathcal{P} = \mathcal{P}_{\text{top-}k} \cup \mathcal{P}_{\text{MST}} \cup \mathcal{P}_{\text{extra}}$.

3.2 First-Order SfM Optimization

Following FastMap [20], we adopt a first-order SfM approach whose per-step cost is independent of 3D point count. Camera intrinsics are estimated via hierarchical interval search using a one-parameter division distortion model; focal length

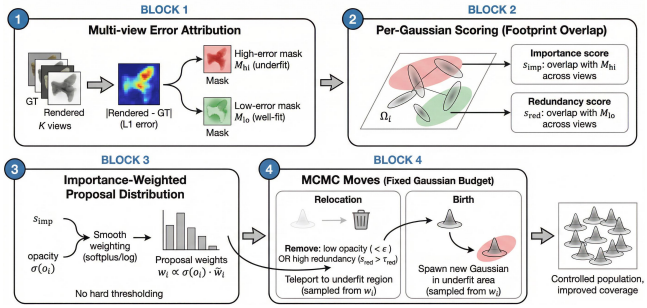

Figure 4: We aggregate multi-view reconstruction errors into per-Gaussian importance and redundancy scores, then use the resulting importance-weighted distribution for both relocation and birth. This reallocates model capacity toward persistently underfit regions while preserving the underlying MCMC population-management framework.

is recovered by maximizing the singular value ratio of the essential matrix. Global rotations are solved by minimizing geodesic distances on SO(3) using a continuous 6D parameterization, and translations are recovered via direction-only consistency with multiple random initializations.

Epipolar Adjustment. The key step is structureless pose refinement using precomputed quadratic forms $W_n \in \mathbb{R}^{9\times 9}$ from epipolar constraints:

$$\mathcal{L}_e = \frac{2}{Z} \sum_{n=1}^{|\mathcal{P}|} \mathbf{e}_n^\top W_n \mathbf{e}_n, \tag{1}$$

where $\mathbf{e}_n = \mathrm{vec}(E_n)$ and Z is a normalization constant. This enables BA-quality refinement with cost independent of point count, implemented with fused CUDA kernels.

3.3 Importance-Guided MCMC Gaussian Allocation

We adopt an MCMC-based approach to Gaussian population management (Figure 4), which differs fundamentally from standard 3DGS adaptive density control (ADC). While ADC uses gradient-based clone/split operations for densification and opacity thresholding for pruning, MCMC-based methods maintain a fixed or slowly-growing Gaussian budget through two operations: **relocation** (teleporting low-value Gaussians to new positions) and **birth** (adding new Gaussians by sampling from existing ones). We extend this framework with importance and redundancy scores derived from multi-view reconstruction error, and use them to define a smooth, importance-weighted sampling distribution that biases both operations toward underfit regions.

Multi-view Error Attribution. Periodically during training, we sample K views and compute per-pixel L1 error maps:

$$e^j_{u,v} = \frac{1}{D} \sum_{d=1}^{D} |r^{j,d}_{u,v} - g^{j,d}_{u,v}|, \tag{2}$$

where r^j and g^j are rendered and ground-truth images for view j, and D=3 is the RGB channel dimension. To make the signal robust across views and training stages, we avoid per-view min-max (which can make "high-error" non-selective late in training) and instead use *robust quantile normalization*:

$$\mathcal{M}^j = \mathrm{clip}\left(\frac{e^j - Q_\ell(e^j)}{Q_h(e^j) - Q_\ell(e^j)}, 0, 1 \right), \tag{3}$$

where Q_ℓ and Q_h denote low/high quantiles (e.g., 5% and 90%). We then define per-view high/low thresholds by quantiles of $\mathcal{M}^j$:

$$\tau^j_{\mathrm{hi}} = Q_{q_{\mathrm{hi}}}(\mathcal{M}^j), \quad \tau^j_{\mathrm{lo}} = Q_{q_{\mathrm{lo}}}(\mathcal{M}^j), \quad (q_{\mathrm{hi}} > q_{\mathrm{lo}}). \tag{4}$$

Importance Score (Underfit Attribution). For each Gaussian $\mathcal{G}_i$ with 2D footprint Ω^j_i in view j, we define a normalized underfit weight:

$$\phi^j_{\mathrm{hi}}(p) = \frac{\max(0, \mathcal{M}^j(p) - \tau^j_{\mathrm{hi}})}{1 - \tau^j_{\mathrm{hi}}}. \tag{5}$$

We aggregate over the footprint and normalize by footprint size to reduce bias toward large projected Gaussians:

$$s^i_{\mathrm{imp}} = \frac{100}{K} \sum_{j=1}^{K} \frac{1}{|\Omega^j_i|} \sum_{p\in\Omega^j_i} \phi^j_{\mathrm{hi}}(p). \tag{6}$$

We report s^i_{imp} on a 0–100 scale (percentage-like), so τ_{imp} is interpretable as the amount of persistent underfit mass required to prioritize a Gaussian. In practice, we estimate $|\Omega^j_i|$ using the projected radii from the renderer (cheap), rather than explicitly enumerating all pixels in Ω^j_i.

Redundancy Score (Well-fit Attribution). We measure well-fit coverage analogously using a normalized low-error weight:

$$\phi^j_{\mathrm{lo}}(p) = \frac{\max(0, \tau^j_{\mathrm{lo}} - \mathcal{M}^j(p))}{\tau^j_{\mathrm{lo}}}. \tag{7}$$

$$s^i_{\mathrm{red}} = \frac{1}{K} \sum_{j=1}^{K} \frac{1}{|\Omega^j_i|} \sum_{p\in\Omega^j_i} \phi^j_{\mathrm{lo}}(p). \tag{8}$$

We min-max normalize s^i_{red} to $[0, 1]$ for thresholding. Gaussians with high s^i_{red} are redundant from a multi-view perspective—they occupy capacity in already well-reconstructed regions.

Importance-Weighted Sampling Distribution. We convert importance into a smooth sampling weight:

$$\tilde{w}_i = \mathrm{softplus}\left(\frac{s^i_{\mathrm{imp}} - \tau_{\mathrm{imp}}}{\tau_{\mathrm{imp}}} \right), \qquad w_i \propto ((1 - \lambda_{\mathrm{mix}})\sigma(o_i) + \lambda_{\mathrm{mix}})\, \tilde{w}_i. \tag{9}$$

where we normalize $\tilde{w}_i$ to have unit mean. The opacity mixing term λ_{mix} prevents under-sampling Gaussians that are

underfit but initially low-opacity (weak coverage), improving exploration beyond "refine only already-visible" regions. This replaces hard thresholding with a differentiable, scale-free weighting around τ_{imp}.

Importance-Guided MCMC Strategy. Our strategy extends the MCMC framework with an importance-weighted sampling distribution shared by both birth and relocation:

- **Relocation**: Gaussians with low opacity ($< \epsilon$) or high redundancy score ($s^i_{\text{red}} > \tau_{\text{red}}$) are reassigned by resampling targets from w_i, recycling capacity from well-fit regions to underfit regions.
- **Birth**: New Gaussians are spawned by sampling parents from w_i, allocating new capacity to consistently underfit regions without hard thresholding.

Unlike standard ADC which can lead to unbounded Gaussian growth, this MCMC-based approach maintains controlled population while the importance guidance ensures capacity is directed toward multi-view consistent reconstruction.

This importance guidance replaces the uniform opacity-weighted sampling of 3DGS-MCMC with w_i; it is a heuristic allocation strategy—not a Metropolis–Hastings proposal—that leaves the underlying SGLD parameter updates unchanged and makes no additional convergence claims. Empirically, it substantially improves sample efficiency under fixed budgets (Table 4).

3.4 Joint Pose Optimization

While first-order SfM provides good initial poses, we jointly refine them during 3DGS training using both photometric and geometric losses.

Photometric Loss. The primary supervision comes from image reconstruction:

$$\mathcal{L}_{\text{photo}} = (1-\lambda_s)\|\hat{I}(\mathbf{x}) - I(\mathbf{x})\|_1 + \lambda_s(1 - \text{SSIM}(\hat{I}, I)), \quad (10)$$

where I and $\hat{I}$ are the ground-truth and rendered images, respectively, and SSIM denotes the Structural Similarity Index Measure. Camera poses $\{T_i\}$ are parameterized with learnable adjustments and optimized jointly with Gaussian parameters.

Bundle Adjustment Loss. To anchor poses to geometric consistency, we incorporate a reprojection-based BA loss:

$$\mathcal{L}_{\text{BA}} = \frac{1}{|\mathcal{O}|} \sum_{(i,j,k)\in\mathcal{O}} \|\pi(T_i, \mathbf{X}_k) - \mathbf{x}_{i,k}\|_2, \quad (11)$$

where $\mathcal{O}$ denotes the set of 2D observations, $\mathbf{X}_k$ are triangulated 3D track points, π is the projection function, and $\mathbf{x}_{i,k}$ is the observed 2D keypoint. The track points and their 2D associations are fixed at SfM initialization; only the camera poses and Gaussian parameters are optimized during joint training. This regularization prevents pose drift by maintaining consistency with sparse geometric constraints from SfM.

Combined Objective. The total training loss combines all terms:

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \lambda_{\text{BA}}\mathcal{L}_{\text{BA}} + \mathcal{L}_{\text{reg}}, \quad (12)$$

where $\mathcal{L}_{\text{reg}}$ includes standard 3DGS regularizers (scale, opacity). Camera 0 is fixed to resolve gauge ambiguity. This joint optimization allows the 3DGS representation and camera poses to co-evolve, leveraging photometric gradients for sub-pixel pose refinement while geometric constraints prevent degenerate solutions.

4 Experimental Evaluation

4.1 Implementation Details

We evaluate on three standard benchmarks: Mip-NeRF 360 [4] (9 scenes), Deep Blending [11] (2 scenes), and Tanks & Temples [18] (2 scenes). We report scene-averaged PSNR, SSIM, and LPIPS [42], end-to-end time (feature extraction, matching, SfM, and training), and Gaussian count N_{GS}. Scene averaging matches the aggregation used for the comparison methods; pooled per-image metrics are reported separately in the supplementary material and are not mixed into Table 1. To summarize cross-dataset behavior, we additionally macro-average the three dataset-level entries, giving each benchmark equal weight. Failed scenes remain in their dataset aggregate; in particular, the *drjohnson* failures of GloSplat-A and VGGT-X are not dropped. The released-code verification uses seed 42 and a single NVIDIA RTX PRO 6000 Blackwell GPU. For Fisher Vector encoding, we train a GMM with M=64 components on SIFT descriptors and retrieve the top-k=20 candidates per image. The Gaussian budget cap is set to N_{GS}=1.5M. For importance-guided MCMC, we use robust normalization quantiles (ℓ, h)=$(0.05, 0.90)$, high/low selection quantiles $(q_{\text{hi}}, q_{\text{lo}})$=$(0.9, 0.1)$, importance threshold τ_{imp}=5 (on a 0–100 importance scale), redundancy threshold τ_{red}=0.9, opacity mixing λ_{mix}=0.05, and K=10 views for score computation. Joint training runs for exactly 30K iterations with λ_{BA}=0.01 and λ_s=0.2. Importance scores are first computed after a pose warmup of 3K iterations and recomputed every T=500 iterations thereafter.

4.2 Main Results

Table 1 compares SalientGS against seven methods that rely on COLMAP preprocessing (upper block) and two concurrent pose-optimizing methods, GloSplat-A [41] and VGGT-X [36] (lower block). All reported times include COLMAP preprocessing for prior methods to reflect true end-to-end cost.

Quality. SalientGS leads the three-benchmark macro-average in PSNR (27.65 dB), SSIM (0.876), and LPIPS (0.147). It is also the only method in Table 1 with top-two LPIPS on every benchmark. On Mip-NeRF 360 it obtains 28.82 dB, 0.853 SSIM, and 0.148 LPIPS with 1.5M Gaussians; its LPIPS is 20.4% below 3DGS-MCMC and 31.5% below FastGS-big. On Deep Blending, it records 29.49 dB / 0.906 / 0.183 and reconstructs *drjohnson*, where GloSplat-A

Table 1: Quantitative comparisons. Times include COLMAP preprocessing for prior methods. Best, 2nd, 3rd. Time in minutes.

Method	Mip-NeRF 360 [4]					Deep Blending [11]					Tanks & Temples [18]				
	Time↓	PSNR↑	SSIM↑	LPIPS↓	N_{GS} ↓	Time↓	PSNR↑	SSIM↑	LPIPS↓	N_{GS} ↓	Time↓	PSNR↑	SSIM↑	LPIPS↓	N_{GS} ↓
w/o Pose Optimization															
3DGS [16]	31.93	27.53	0.812	0.221	2.63M	30.77	29.71	0.903	0.241	2.46M	22.34	23.71	0.850	0.170	1.57M
3DGS-MCMC [17]	32.41	28.01	0.835	0.186	3.23M	31.25	29.78	**0.912**	0.237	2.95M	22.89	24.40	**0.869**	0.149	1.85M
Mini-Splatting [6]	28.69	27.32	0.821	0.217	0.53M	24.35	29.99	0.907	0.244	0.56M	20.06	23.46	0.844	0.181	0.30M
Speedy-splat [9]	24.38	26.91	0.781	0.295	**0.30M**	21.75	29.42	0.898	0.272	**0.25M**	17.32	23.38	0.816	0.242	**0.18M**
Taming-3DGS [22]	16.36	27.48	0.794	0.261	0.68M	14.06	29.50	0.894	0.278	0.29M	13.71	23.89	0.833	0.214	0.32M
DashGaussian [5]	17.35	27.73	0.817	0.218	2.40M	15.16	29.65	0.906	0.246	1.94M	15.28	24.00	0.853	0.178	1.21M
FastGS-big [29]	14.58	27.93	0.820	0.216	1.15M	13.00	**30.12**	0.907	0.243	0.65M	13.03	24.39	0.855	0.175	0.54M
w/ Pose Optimization															
GloSplat-A [41]	22.00	**28.86**	**0.862**	**0.139**	3.00M	19.15	18.45*	0.583*	0.508*	3.00M	24.87	22.15	0.805	0.147	3.00M
VGGT-X [36]	73.71	26.49	0.782	0.177	3.00M	57.24	18.25†	0.622†	0.545†	3.00M	61.12	23.05	0.818	0.138	3.00M
SalientGS (Ours)	**11.79**	28.82	0.853	0.148	1.50M	**10.04**	29.49	0.906	**0.183**	1.50M	**10.03**	**24.65**	**0.869**	**0.109**	1.50M

*GloSplat-A and †VGGT-X consistently fail on the *drjohnson* scene, significantly degrading their averaged Deep Blending metrics.

Table 2: Component ablation on Mip-NeRF 360. ΔPSNR is relative to the full configuration.

Configuration	PSNR↑	SSIM↑	LPIPS↓	ΔPSNR
(A) Importance-Guided MCMC				
Full model	**28.82**	**0.853**	**0.148**	—
w/o Guided Birth	28.72	0.847	0.153	−0.10
w/o Guided Reloc.	28.77	0.849	0.149	−0.05
w/o Both (Vanilla MCMC)	28.72	0.848	0.149	−0.10
w/o Footprint Normalization	22.70	0.684	0.422	−6.12
(B) Densification Strategy				
Standard ADC (clone/split)	27.67	0.828	0.162	−1.15

Table 3: Pose/SfM ablation on Mip-NeRF 360. ΔPSNR is relative to the full configuration.

Configuration	PSNR↑	SSIM↑	LPIPS↓	ΔPSNR
(A) Joint Pose Optimization				
Full model	**28.82**	**0.853**	**0.148**	—
w/o BA Loss (photometric-only)	28.70	0.845	0.156	−0.12
Freeze Poses after SfM Init	28.32	0.836	0.160	−0.50
(B) Sensitivity to SfM Initialization Quality				
Higher k=40 retrieval	**28.86**	**0.860**	**0.134**	+0.04
Full model (k=20 retrieval)	28.82	0.853	0.148	—
Lower k=10 retrieval	27.35	0.833	0.169	−1.47
Lower k=5 retrieval†	26.57	0.774	0.257	−2.25

†At k=5, 1/9 scenes fails catastrophically (stump: 14.97 dB PSNR), significantly degrading the average.

and VGGT-X fail. On Tanks & Temples, it leads PSNR (24.65) and LPIPS (0.109) and ties the best SSIM (0.869). The macro-average thus measures consistency rather than one favorable dataset.

Efficiency and Model Size. Including every front-end and training stage, SalientGS averages 11.79, 10.04, and 10.03 minutes on the three benchmarks, respectively. It is the fastest end-to-end method in each block of Table 1; on Mip-NeRF 360 it is 1.87× faster than GloSplat-A and 6.25× faster than VGGT-X, with half their Gaussian count. The released CSV and runner make this hardware-dependent comparison reproducible.

4.3 Ablation Study

We ablate each key component of SalientGS on the Mip-NeRF 360 benchmark to quantify individual contributions. All variants share the same first-order SfM initialization and Fisher Vector matching; only the 3DGS training stage differs unless otherwise noted. We use the corrected full-model aggregate, 28.82 dB / 0.853 / 0.148, consistently in the main comparison and ablations; all deltas below are recomputed against this value.

Importance-Guided MCMC (Group A). Against vanilla MCMC, guidance adds 0.10 dB and lowers LPIPS from 0.149 to 0.148. Birth guidance provides the larger PSNR contribution; relocation is most useful when birth is also guided. Removing footprint normalization costs 6.12 dB because large projected primitives otherwise dominate the scores.

Densification Strategy (Group B). Replacing guided MCMC management with standard ADC costs 1.15 dB, supporting multi-view redistribution under a fixed budget.

Joint Pose Optimization and SfM Initialization. Table 3 presents ablations on joint optimization and SfM initialization quality.

Joint Pose Optimization (Group A). Removing BA costs 0.12 dB, whereas freezing poses costs 0.50 dB. Thus photometric refinement provides most of the recovery and reprojection anchoring adds a further measurable gain.

SfM Initialization Quality (Group B). Increasing retrieval from k=20 to 40 gives a small 0.04 dB gain, whereas reducing it to 10 and 5 costs 1.47 and 2.25 dB, respectively; at k=5, *stump* fails at 14.97 dB. Joint optimization therefore cannot replace a connected, reliable view graph.

Gaussian Budget Efficiency. Guidance improves every tested budget, with the gain generally decreasing from +0.27 dB at 500K to +0.09 dB at 3M. Guided 1M (28.81 dB)

Table 4: Gaussian budget efficiency on Mip-NeRF 360. PSNR at varying N_{GS} caps for vanilla MCMC vs. importance-guided MCMC. Δ shows the gain from importance guidance.

N_{GS} Cap	Vanilla MCMC	Guided MCMC	ΔPSNR
500K	28.19	28.46	+0.27
1.0M	28.59	28.81	+0.22
1.5M	28.72	28.82	+0.10
2.0M	28.80	28.93	+0.13
3.0M	28.88	**28.97**	+0.09

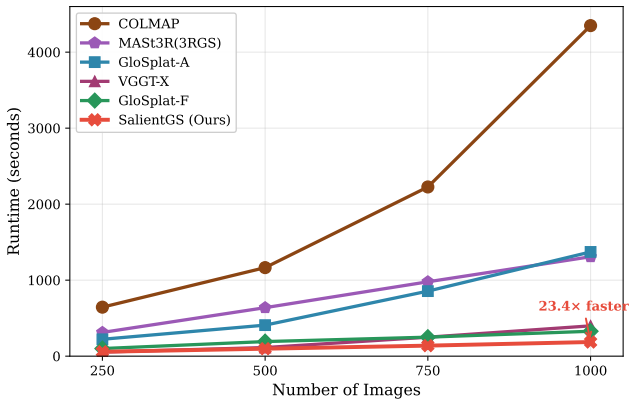

Figure 5: SfM runtime vs. image count. SalientGS achieves up to 23× speedup over COLMAP with near-linear scaling.

already exceeds vanilla 1.5M (28.72 dB), showing the largest value under tight capacity.

Schedule and normalization sensitivity. The supplementary material reports every schedule and normalization setting. Nearby alternatives can shift individual metrics slightly, while the released model uses one fixed default across all scenes.

4.4 SfM Runtime Comparison

Figure 5 compares SfM runtime on the Courthouse scene (250–1000 images) against COLMAP, MASt3R [19], VGGT [36], and GloSplat [41]. SalientGS achieves 13.5–23.4× speedup over COLMAP with near-linear scaling; the full numeric breakdown is provided in the Supplementary "SfM Runtime Details" section.

4.5 Qualitative Results

We present qualitative comparisons on representative Mip-NeRF 360 scenes in Figure 6. For each view, we show the full rendered image alongside a zoomed crop of the region with the greatest perceptual difference (highest LPIPS gap) between our method and the baselines.

Success Cases. SalientGS shows the strongest advantages on outdoor scenes with complex, high-frequency textures.

Table 5: Pose evaluation on ETH3D SLAM [31]. Three SfM initializations—FastMap [20], COLMAP [30], GLOMAP [27]—evaluated before and after joint optimization (§3.4). Results on two subsets: 39 sequences where SfM and joint training both succeed, and 34 where 3DGS reconstruction also succeeds (PSNR>20).

Subset	Method	Recall@ 0.1m	AUC@ 0.1m	AUC@ 0.5m
Train-succ. (39 seq.)	FastMap (init)	50.5	37.2	58.2
	FastMap + joint opt	**57.0**	**42.8**	**61.0**
	COLMAP (init)	50.2	38.5	59.8
	COLMAP + joint opt	52.0	40.5	60.4
	GLOMAP (init)	55.5	40.1	60.2
	GLOMAP + joint opt	56.5	42.3	60.7
Recon. succ. (34 seq.)	FastMap (init)	55.8	41.0	61.2
	FastMap + joint opt	**62.0**	**47.5**	**64.5**
	COLMAP (init)	61.0	40.8	60.3
	COLMAP + joint opt	61.2	42.5	61.0
	GLOMAP (init)	60.2	44.5	63.5
	GLOMAP + joint opt	61.4	47.0	64.1

On *garden* and *stump*, importance-guided MCMC allocates Gaussians to resolve fine foliage and bark textures that Vanilla MCMC renders as blurred masses. On *counter*, SalientGS recovers sharper edges on kitchen objects that other methods over-smooth. VGGT-X consistently exhibits ghosting and geometric distortions in outdoor scenes due to feed-forward pose estimation errors.

Failure Cases. On indoor scenes with uniform textures (*room*, *kitchen*), VGGT-X's dense feed-forward initialization provides stronger geometric constraints than our first-order SfM, reducing the benefit of importance-guided allocation. Detailed per-scene LPIPS comparisons are provided in the Supplementary "Per-Image LPIPS Analysis" section.

4.6 Pose Evaluation on ETH3D SLAM

To evaluate pose accuracy independently of rendering quality, we benchmark on 45 training sequences from the ETH3D SLAM dataset [31], which provides millimeter-accurate ground-truth poses from motion capture. These sequences span 17 scene groups covering sparse textures, dynamic objects, and drastic illumination changes. Following the GLOMAP [27] evaluation protocol, we report absolute camera position accuracy after robust Procrustes alignment: **Recall@0.1m** (fraction of cameras within 10 cm of ground truth), and **AUC** at 0.1 m and 0.5 m thresholds.

This experiment is designed to isolate the contribution of joint optimization of pose and appearance from SfM initialization quality. We evaluate three SfM methods—**FastMap** [20], **COLMAP** [30], and **GLOMAP** [27]—each in two configurations: (1) poses used directly without refinement (init), and (2) poses after joint training with photometric and BA losses (§3.4). This cross-product design reveals both the effect of initialization quality and the consistent benefit of joint refinement.

Figure 6: Qualitative comparison on Mip-NeRF 360. The first three panels show success cases where SalientGS achieves large LPIPS improvements over baselines. The final panel shows the *room* failure case, where SalientGS (LPIPS = 0.312) underperforms VGGT-X (LPIPS = 0.244) on an indoor scene with uniform textures.

Table 5 presents the results. Of our 45 sequences, SfM successfully registers cameras in 43 (96%); 39 of these also complete joint training (Train-succ.), and 34 achieve full reconstruction success (Recon. succ., PSNR>20).

Joint optimization consistently improves all pose metrics regardless of SfM initialization. The improvement is largest for FastMap, the weakest initialization: AUC@0.1m improves by +5.6 (39 seq.) and +6.5 (34 seq.), compared to +2.0/+1.7 for COLMAP and +2.2/+2.5 for GLOMAP. Strikingly, FastMap + joint optimization achieves the best pose accuracy across all metrics on both subsets, surpassing even GLOMAP + joint optimization (e.g., AUC@0.1m 42.8 vs. 42.3 on 39 seq.; 47.5 vs. 47.0 on 34 seq.). We attribute this to the nature of first-order SfM errors: they are smooth and systematic rather than discrete outliers, making them particularly amenable to correction via photometric and reprojection gradients. This result validates the SalientGS pipeline design—fast first-order SfM initialization paired with joint refinement not only matches but slightly exceeds the pose accuracy of stronger SfM baselines.

5 Conclusion and Limitations

We presented SalientGS, a unified SfM-to-3DGS pipeline for fast reconstruction with robust quality under a fixed Gaussian budget. In the released-code 13-scene verification, SalientGS achieves the best three-benchmark macro-average PSNR, SSIM, LPIPS, and end-to-end runtime among all methods in the main comparison. This overall result is driven by consistency: SalientGS remains strong across all three datasets, whereas competing pose-optimizing pipelines suffer severe failures on Deep Blending. The result supports the practical value of combining retrieval-based matching, first-order SfM, joint pose refinement, and importance-guided allocation. At the same time, individual dataset metrics show room for improvement, particularly Deep Blending PSNR/SSIM and fine-detail LPIPS on several Mip-NeRF 360 scenes.

Design Choice: Minimal Per-Scene Tuning. SalientGS uses a single fixed set of hyperparameters across all scenes, unlike methods that use per-scene schedules [17, 29]. This simplifies deployment and makes the released runner deterministic under a fixed seed, but it can leave scene-specific quality on the table (see the Supplementary "Threshold Sensitivity Sweep" section).

Limitation: Pose Optimization vs. Speed. A fundamental tension exists between camera pose optimization and training efficiency. The verified end-to-end runtime is competitive, but joint optimization still adds training-stage overhead relative to a frozen-pose run on identical hardware. Reducing that overhead without weakening geometric correction remains open.

Limitation: Upstream Component Dependencies. Joint optimization is effective only when upstream SfM succeeds: of 45 ETH3D SLAM sequences, 2 fail SfM entirely and 4 fail joint training, though the remaining 39 show consistent improvement (AUC@0.1m gains in 67% of scenes). Additionally, Fisher Vector retrieval may fail under severe appearance variation where learned descriptors would provide more robust pair selection, and first-order SfM is sensitive to sparse coverage and degenerate camera motions such as collinear configurations. These limitations highlight that our joint optimization refines rather than substitutes for successful SfM initialization.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 5297–5307.

[2] Gabriele Barbarani, Gabriele Trivigno, Carlo Berton, Debora Mereu, and Carlo Masone. 2024. MegaLoc: One Retrieval to Place Them All. In *European Conference on Computer Vision*. Springer, Cham, 382–400.

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021.

Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Piscataway, NJ, USA, 5855–5864.

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 5470–5479.

[5] Youyu Chen, Junjun Jiang, Kui Jiang, Xiao Tang, Zhihao Li, Xianming Liu, and Yinyu Nie. 2025. DashGaussian: Optimizing 3D Gaussian Splatting in 200 Seconds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, Piscataway, NJ, USA, 11146–11155.

[6] Guangchi Fang and Bing Wang. 2024. Mini-splatting: Representing scenes with a constrained number of gaussians. In *European Conference on Computer Vision*. Springer, Cham, 165–181.

[7] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. 2024. COLMAP-Free 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 20796–20805.

[8] Venu Madhav Govindu. 2001. Combining two-view constraints for motion estimation. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2. IEEE, Piscataway, NJ, USA, II–II.

[9] Alex Hanson, Allen Tu, Geng Lin, Vasu Singla, Matthias Zwicker, and Tom Goldstein. 2025. Speedy-splat: Fast 3d gaussian splatting with sparse pixels and sparse primitives. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, Piscataway, NJ, USA, 21537–21546.

[10] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. 2013. Rotation averaging. *International Journal of Computer Vision* 103, 3 (2013), 267–305.

[11] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. 2018. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)* 37, 6 (2018), 1–15.

[12] Boyao Huang et al. 2025. 3R-GS: Efficient Gaussian Splatting without SfM Using 3D Reconstruction Model. *arXiv preprint arXiv:2503.04946* (2025).

[13] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 3304–3311.

[14] Zihao Ji et al. 2025. Hierarchical Tracking 3DGS: Robust SfM-Free 3D Gaussian Splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, Piscataway, NJ, USA, 6170–6180.

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4, Article 139 (July 2023), 14 pages. https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[17] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Jeff Tseng, Hossam Isber, and Andrea Tagliasacchi. 2024. 3D Gaussian Splatting as Markov Chain Monte Carlo. In *Advances in Neural Information Processing Systems*, Vol. 37. Curran Associates, Inc., Red Hook, NY, USA.

[18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[19] Vincent Leroy, Yohann Cabon, and Jerome Revaud. 2024. Grounding Image Matching in 3D with MASt3R. In *European Conference on Computer Vision*. Springer, Cham, 71–91.

[20] Jiahao Li, Haochen Wang, Muhammad Zubair Irshad, Igor Vasiljevic, Matthew R Walter, Vitor Campagnolo Guizilini, and Greg Shakhnarovich. 2026. FastMap: Revisiting Structure from Motion through First-Order Optimization. In *Proceedings of the International Conference on 3D Vision*. IEEE, Piscataway, NJ, USA.

[21] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. 2021. BARF: Bundle-Adjusting Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Piscataway, NJ, USA, 5741–5751.

[22] Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Markus Steinberger, Francisco Vicente Carrasco, and Fernando De La Torre. 2024. Taming 3dgs: High-quality radiance fields with limited resources. In *SIGGRAPH Asia 2024 Conference Papers*. ACM, New York, NY, USA, 1–11.

[23] Daniel Martinec and Tomas Pajdla. 2007. Robust rotation and translation estimation in multiview reconstruction. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 1–8.

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*. Springer, Cham, 405–421.

[25] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. 2016. OpenMVG: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*. Springer, Cham, 60–74.

[26] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. In *ACM SIGGRAPH 2022 Conference Proceedings*. ACM, New York, NY, USA, 1–15.

[27] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes L Schönberger. 2024. Global structure-from-motion revisited. In *European Conference on Computer Vision*. Springer, Cham, 58–77.

[28] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*. Springer, Cham, 143–156.

[29] Shiwei Ren, Tianci Wen, Yongchun Fang, and Biao Lu. 2025. FastGS: Training 3D Gaussian Splatting in 100 Seconds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. IEEE, Piscataway, NJ, USA, 18540–18550.

[30] Johannes L Schönberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 4104–4113.

[31] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. 2019. Bad SLAM: Bundle Adjusted Direct Visual SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 134–144.

[32] Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2. IEEE, Piscataway, NJ, USA, 1470–1477.

[33] Cameron Smith, Lily Goli, David Chen, Adam W Harley, and Leonidas Guibas. 2024. FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent. In *European Conference on Computer Vision*. Springer, Cham, 345–363.

[34] Chris Sweeney. 2015. Theia: A fast and scalable structure-from-motion library. In *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, New York, NY, USA, 693–696.

[35] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. 2023. SPARF: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 4190–4200.

[36] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Tagliasacchi, David Novotny, Christian Rupprecht, et al. 2025. VGGT: Visual Geometry Grounded Transformer. *arXiv preprint arXiv:2503.11651* (2025).

[37] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. 2024. VGGSfM: Visual Geometry Grounded Deep Structure From Motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 21686–21697.

[38] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 20697–20709.

[39] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. 2021. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064* (2021).

[40] Kyle Wilson, David Bindel, and Noah Snavely. 2020. On the distribution of minima in intrinsic-metric rotation averaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 11997–12006.

[41] Tianyu Xiong, Rui Li, Linjie Li, and Jiaqi Yang. 2026. GloSplat: Joint Pose-Appearance Optimization for Faster and More Accurate 3D Reconstruction. *arXiv preprint arXiv:2603.04847* (2026). https://arxiv.org/abs/2603.04847

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 586–595.