FastSplatting: SfM free Ultra Fast Gaussian Splatting.

Anurag Dalal[1], Daniel Hagen[1], Kjell G. Robbersmyr[1], and Kristian Muri Knausgård[1]

***Abstract*— 3D Gaussian Splatting (3DGS) has recently emerged as a powerful technique for high-quality and computationally efficient 3D scene reconstruction. Alongside 3D Foundation Models (3DFMs) have also gained tremendous popularity, enabling applications ranging from VR and robotics to autonomous navigation. However, most existing 3DGS pipelines depend on Structure-from-Motion (SfM) to estimate camera poses before reconstruction. This step is slow and computationally heavy, with the complexity rising exponentially with the number of input images, and often limits the use of 3DGS in real-time or large-scale scenarios. In this paper, we introduce a new method for generating fast and accurate 3DGS reconstructions without using traditional SfM. This method utilizes 3DFM for initialization of the camera poses and the point cloud for gaussian initialization also jointly optimize both camera poses and gaussian primitives within the 3DGS framework, allowing the system to quickly converge even from rough initialization. Our approach uses refined optimization strategies and efficient loss function to achieve high-quality reconstruction with as few as 50-60 input views. Through extensive experiments across multiple datasets, it is shown that this method produces ready to use 3DGS models at a fraction of the time required by traditional pipelines, making it suitable for real-time and dynamic environments.**

I. Introduction

3DGS [1] has transformed neural rendering through its efficiency and quality. However, similar to many novel view synthesis techniques, it relies significantly on precise camera poses obtained from SfM systems utilizing COLMAP [2], [3]. SfM is computationally intensive and can be unreliable in challenging scenarios, such as texture-less scenes or limited viewpoints [4]. This dependency on SfM not only increases the overall processing time but also any potential inaccuracies in camera pose estimation, adversely affect the quality of 3D reconstruction.

Recent advancements in 3DFMs [5]–[7] have shown promise in reducing the reliance on SfM by employing alternative methods for camera pose estimation and scene reconstruction. These methods often utilize transformers based deep learning techniques to infer camera poses directly from multi-view images of a scene. By integrating these approaches with 3DGS, it is possible to achieve rapid initialization of the gaussian primitives in order of milliseconds and accurate 3D reconstructions without the overhead of traditional SfM pipelines, which takes minutes. VGGT-X [8], which is based on the 3DFM VGGT [5] is one such method that has demonstrated the ability to estimate camera poses along with a point cloud with improved accuracy and low latency making it a suitable candidate for integration with 3DGS.

Another technique that has shown potential in this domain is 3R-GS [9], which optimizes camera poses along with 3DGS. By jointly optimizing both the camera parameters and the Gaussian primitives, 3R-GS can achieve high-quality reconstructions even with a limited number of viewpoints. This approach not only reduces the dependency on accurate initial camera poses but also enhances the overall robustness of the reconstruction process.

This method combines the strengths of VGGT-X and 3R-GS to create a novel pipeline for ultra-fast Gaussian splatting-based 3DGS without relying on SfM. By leveraging VGGT-X for initial camera pose estimation and subsequently refining these poses using the joint optimization framework of 3R-GS, thereby achieving rapid and accurate 3D reconstructions with few as viewpoints, also being robust enough to cater for hundreds of views. This approach not only accelerates the reconstruction process but also improves the quality of resulting 3DGS, making it suitable for real-time applications in various fields such as virtual reality, robotics, and augmented reality.

Furthermore, to address the challenges of limited viewpoints, a depth-guided loss function is introduced that leverages geometric priors to enhance reconstruction quality. When fewer viewpoints are available, traditional photometric losses may be insufficient to constrain the optimization process, leading to ambiguous or inaccurate reconstructions. Our depth-guided loss incorporates depth information from the 3DFM to provide additional geometric constraints during the joint optimization of camera poses and Gaussian parameters. This approach helps maintain geometric consistency and improves the fidelity of rendered images, particularly in regions with sparse view coverage. By combining photometric and geometric supervision, our method achieves much faster reconstruction. In summary, our contributions are as follows:

- A novel pipeline that achieve ultra-fast 3DGS without the need for SfM.
- Optimizing camera poses and 3DGS jointly to enhance reconstruction quality.
- Unique loss function and optimization strategy tailored for low-viewpoint scenarios.

II. Related Work

This section reviews the relevant literature on 3DGS and 3DFMs, focusing on their methodologies, applications, and

*We would like to extend our sincere thanks to Aust Agder utviklings og kompetansefond (AAUKF) for the generous funding of the Arven etter Dannevig (The Legacy of Dannevig) project, nr 62/22 which has been instrumental in the completion of this paper.

[1]Department of Engineering Sciences, University of Agder, Jon Lilletuns vei 9, 4879 Grimstad, Norway.
anurag.dalal@uia.no, daniel.hagen@uia.no, kjell.g.robbersmyr@uia.no, kristianmk@ieee.org

advancements in the field.

1) 3DGS: 3D Gaussian Splatting (3DGS) [1] represents a scene as a set of anisotropic Gaussian primitives placed directly in 3D space rather than on a fixed voxel or mesh structure. Each Gaussian stores a mean (3D position), covariance (shape/orientation), color (or SH coefficients), and opacity. Rendering proceeds via forward GPU rasterization, gaussians are projected to screen space and composited front-to-back using alpha blending, enabling real-time novel view synthesis without expensive ray marching. Differentiability of the splat pipeline allows gradient-based optimization of both appearance and geometry from posed images. Compared to NeRF-style volumetric integration, 3DGS achieves large speedups (interactive FPS) while maintaining high fidelity, thanks to adaptive spatial density and learned anisotropy that aligns splats with local surface structure. This efficiency makes 3DGS well suited for our accelerated pipeline where reliable poses and depth from 3DFMs seed initial Gaussian placement and reduce subsequent optimization iterations.

The emergence of 3DGS has rapidly transformed 3D reconstruction workflows across multiple domains by collapsing the time gap between capture and high-quality visualization [10]. In immersive media (VR/AR), its real-time rendering enables interactive scene exploration and on-the-fly relighting with far lower latency than NeRF-style models. In robotics and autonomous navigation, fast splat updates allow dynamic map refinement and uncertainty-aware perception layers that integrate smoothly with planning loops. 3DGS can also be used in cultural heritage digitization, benefiting from efficient reconstruction of complex surfaces (ornamented facades, sculptures) without dense meshing overhead, while remote sensing and for UAV inspection pipelines 3DGS can be used to model fine structural details (vegetation, infrastructure) under constrained compute at the edge. Synthetic data generation and simulation leverage controllable splat attributes (opacity, anisotropy) to produce photo-realistic training environments with rapid iteration. Collectively, these advances shift 3DGS from a pure rendering technique to an enabling geometric substrate that accelerates downstream tasks—segmentation, tracking, depth refinement—by providing both differentiable and real-time access to scene structure.

2) 3DFM: 3DFMs unify multi-view perception tasks like camera pose estimation, depth prediction, point map generation, and correspondence reasoning within large transformer architectures trained on diverse large-scale image/video corpora. Most 3DFMs uses large vision transformers such as DINO [11]–[13] or ViT [14] as the backbone. This work focus on the recent VGGT [5] rather than DUSt3R [6], or MASt3R [7] on an accelerated Gaussian Splatting pipelines because of its faster inference time and better accuracy.

DUSt3R [6] focuses on dense multi-view matching and geometry reconstruction through learned pixel-aligned descriptors and global alignment. It produces relative camera poses and a fused point cloud by aggregating per-view features into a shared 3D latent representation. MASt3R [15] extends multi-view reconstruction with improved attention aggregation over views, yielding refined depth, normal, and point map predictions along with camera geometry. Its multi-view fusion improves completeness and reduces holes relative to single-view depth estimators. For downstream tasks: normal and depth enhance open-world segmentation by improving region separation (geometry-aware prompting); consistent multi-view depth can bootstrap indoor/outdoor scale estimation modules; and predicted surface orientation can modulate splat anisotropy (aligning Gaussian covariance with local tangent planes) for sharper renderings with fewer splats.

VGGT [5] proposes a unified transformer with DINO as its backbone as shown in Figure 1 that ingests a set of unordered multi-view images and jointly infers (a) camera extrinsics/intrinsics, (b) per-pixel depth maps, (c) dense tracking features, and (d) point maps or normalized 3D coordinates. It leverages cross-frame attention to build a geometry-aware latent space without an explicit SfM step. The model learns implicit epipolar constraints through large-scale pretraining and can output camera parameters in a single forward pass. VGGT-X [8] further improves pose accuracy and depth quality by integrating dense feature matching and geometric consistency losses during fine-tuning. Its enhanced pose estimates and depth maps provide reliable initialization for downstream 3D reconstruction tasks, making it well suited for our SfM-free Gaussian splatting pipeline.

In summary, 3DFMs furnish a unified set of geometric primitives—poses, depth, normals, point maps, and confidence—that dramatically compress the traditionally sequential SfM+MVS pipeline into one forward pass, while enabling richer downstream adaptation for depth refinement, semantic segmentation, and for this case accelerated Gaussian Splatting.

III. Methodology

This section details our proposed pipeline that integrates VGGT-X and 3R-GS for ultra-fast Gaussian splatting-based 3DGS without relying on SfM. Finally, the unique loss function and training strategy that is utilized to handle less viewpoints scene as well aids in achieving faster convergence is discussed.

Camera Parameter and Depth Estimation Mechanism: Standard 3D Gaussian Splatting (3DGS) is highly sensitive to initialization. Because Gaussian primitives have limited ability to move far from their starting positions, small pose or point-cloud errors can trap them in poor local minima. Photometric losses provide only local gradients, making it hard for misaligned Gaussians to correct themselves. Additionally, 3DGS's densification strategy relies on hand-tuned gradient thresholds, which becomes unstable when also optimizing camera poses. VGGT processes a batch of images through (1) patch embedding, (2) multi-view cross-attention blocks, and (3) task heads. Cross-attention mixes tokens across frames, allowing the model to infer relative orientation and translation by optimizing consistency of learned geometric tokens. Pose head outputs extrinsics (rotation, translation)

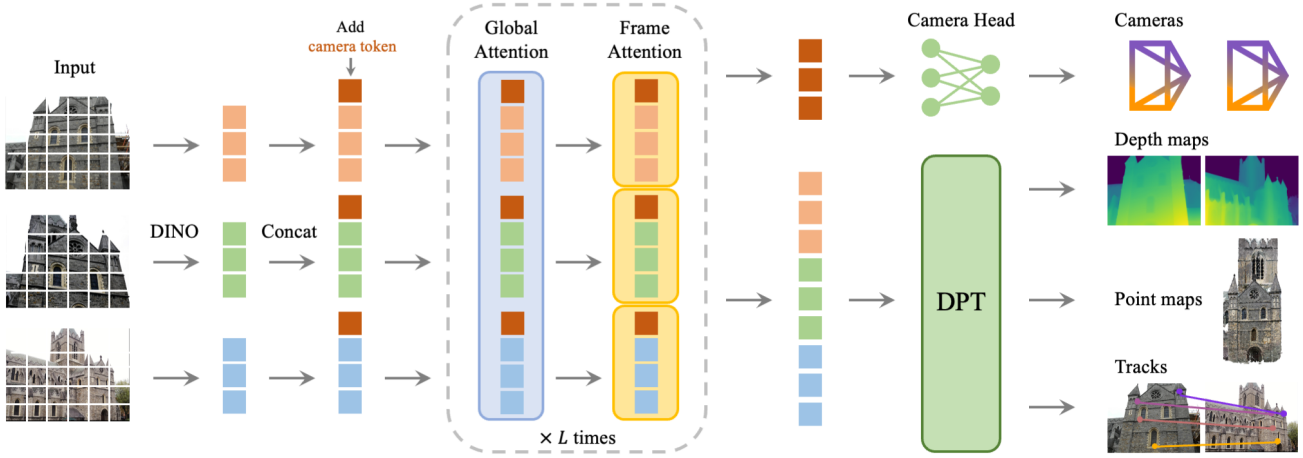

Fig. 1: VGGT architecture illustrating multi-view fusion and task heads. Source: Wang et al. [5]

and intrinsics or focal length. Depth head predicts per-pixel depth aligned to those poses; a point map head may output scene coordinates in a canonical frame. Internally, learned attention patterns approximate epipolar reasoning: correspondences produce constraints the transformer aligns implicitly, akin to neural bundle adjustment without iterative classical optimization. Confidence maps or uncertainty estimates further filter unreliable regions, enabling robust initialization for downstream optimization.

Camera pose optimization: Standard 3D Gaussian Splatting (3DGS) suffers from high sensitivity to initialization because Gaussian primitives have limited ability to move far from their initial positions. When the starting point cloud or camera poses contain errors, the photometric rendering loss provides only local gradients, which makes it difficult for a misaligned Gaussian to escape poor local minima. This often results in sub-optimal convergence and degraded scene reconstruction quality. Furthermore, adaptive density control in 3DGS relies on gradient-magnitude thresholds that require manual tuning, which becomes increasingly unstable when jointly optimizing Gaussian primitives and camera poses.

To address these issues, 3DGS-MCMC strategy is employed [16], which reformulates Gaussian optimization as Markov Chain Monte Carlo sampling. The training process is interpreted as sampling from a distribution $p(G)$ that assigns higher probability to Gaussian configurations that accurately reproduce the training images. Under this formulation, standard 3DGS optimization behaves similarly to Stochastic Gradient Langevin Dynamics (SGLD), with updates of the form

$$G \leftarrow G + a\,\nabla_G \log p(G) + b\,\eta \tag{1}$$

where η is exploration noise and a and b control the balance between convergence and exploration. The injected noise enables Gaussians to escape local minima caused by imperfect initialization. Moreover, 3DGS-MCMC replaces heuristic densification and pruning strategies with principled state transitions, and incorporates a regularizer that encourages parsimonious Gaussian usage. Together, these components allow robust joint optimization of camera poses and Gaussian primitives.

Despite improved robustness, camera poses may still exhibit shared global drift: even if relative poses are accurate, the entire camera set may be shifted or rotated away from the true configuration. Directly optimizing each pose independently ignores these global correlations and can distort otherwise correct relative geometries due to the inherent non-convexity of pose optimization. To address this, an MLP-based global pose refiner is introduce that predicts pose corrections from a learned camera embedding. For each camera i, the correction is given by

$$\Delta T_i = R_{\text{MLP}}(z_i), \tag{2}$$

where z_i is a learnable embedding associated with camera i, and the output consists of translation $\Delta t_i \in \mathbb{R}^3$ and rotation $\Delta r_i \in \mathbb{R}^6$ components. The MLP is initialized with a zero-mean prior to ensure stable and unbiased refinement. By sharing the refinement network across all cameras, the method captures global pose relationships and improves the accuracy of camera pose adjustment, outperforming direct per-camera optimization.

Loss function: 3DGS uses a combination of $\mathcal{L}_1$ loss and a Structural Similarity Index Measure (SSIM) [17] loss, to supervise the rendered images against ground truth. The photometric loss is defined as:

$$\mathcal{L}_{\text{Photo}} = (1 - \lambda_{\text{SSIM}})\mathcal{L}_{\text{L1}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} \tag{3}$$

where $\mathcal{L}_{\text{L1}}$ is the Mean Absolute Error between the rendered image I_{render} and the ground truth image I_{GT}:

$$\mathcal{L}_{\text{L1}} = \frac{1}{N}\sum_i |I_{\text{render}}(i) - I_{\text{GT}}(i)| \tag{4}$$

and $\mathcal{L}_{\text{SSIM}}$ is the Structural Similarity Index Measure loss between I_{render} and I_{GT}.

A depth guided loss term ($\mathcal{L}_{\text{Depth}}$) is also introduced along with the $\mathcal{L}_{\text{Photo}}$. This $\mathcal{L}_{\text{Depth}}$ uses the normalized alpha channel of gaussian splatting render (D_{render}) and the normalized

depth maps from VGGT-X (D_{VGGTX}).

$$\mathcal{L}_{\text{Depth}} = \frac{1}{N} \sum_{i} |D_{\text{render}}(i) - D_{\text{VGGTX}}(i)| \tag{5}$$

The overall total loss, including all optional terms, is expressed as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Photo}} + \lambda_{\text{Depth}} \mathcal{L}_{\text{Depth}} \tag{6}$$

Integration into Our Pipeline. We first pass the multi-view images through VGGT to obtain initial camera parameters and depth. These camera poses and depths serve as high-quality initializations for 3D Gaussian Splatting by: seeding Gaussian centers from back-projected depth, constraining early optimization with reliable intrinsics/extrinsics, and pruning outlier pixels using confidence masks produced by the transformer. Figure 2 illustrates the overall FastSplatting pipeline.

IV. Experiments

A. Datasets

The proposed method is evaluated on several publicly available datasets that are commonly used for 3D reconstruction and novel view synthesis tasks. These datasets offer a variety of scenes, viewpoints, and complexities to comprehensively assess the performance of the method.

- **Mip-NeRF 360** [18] dataset consists of real-world scenes captured with a camera performing a full 360-degree rotation around a central point of interest. Each scene features complex geometry and rich appearance details, including reflective surfaces, foliage, and cluttered backgrounds that challenge view synthesis methods. The dataset provides high-resolution images with consistent exposure settings to minimize photometric variation during capture. By spanning both indoor and outdoor environments, Mip-NeRF 360 offers diverse lighting conditions, occlusions, and depth ranges. Its full-360 capture setup and challenging visual complexity make it a widely used benchmark for evaluating 3D reconstruction, neural radiance field methods, and Gaussian splatting approaches.
- **Tanks and Temples** [19] dataset is a widely adopted benchmark for evaluating real-world 3D reconstruction and novel view synthesis methods. In our experiments, we use scenes from the supplementary database subset, denoted as tandt_db. The tandt_db subset includes challenging indoor environments such as drjohnson, which features museum-like sculptures and complex lighting, and playroom, a cluttered indoor room with diverse objects and occlusions. From the main T&T benchmark, the train scene and the outdoor truck scene is included, the latter being notable for its large-scale geometry, vegetation, and detailed background structures. Together, these scenes provide a diverse set of indoor and outdoor conditions, enabling rigorous evaluation of reconstruction robustness and generalization.
- **RobustNeRF** [20] dataset consists of both natural and synthetic scenes designed to evaluate robustness under varying levels of visual distraction. The natural portion includes seven real-world scenes captured across streets, an apartment, and a robotics lab, where distractor objects are deliberately moved or allowed to move between frames to simulate long-term or uncontrolled capture conditions. These scenes vary widely in complexity, containing anywhere from a single distractor to up to 150 distractors, and include cases with strong view-dependent effects such as Street1, Street2, and Gloss. Frames are captured without temporal ordering, and additional distractor-free images are provided for quantitative evaluation. To complement these real scenes, the dataset also includes synthetic sequences generated with the Kubric engine, where simple geometric objects are placed in a texture-less room, and a subset of them move between frames. By controlling object count, size, and motion, the synthetic setups allow precise analysis of distraction levels and their effect on RobustNeRF's performance.

B. Training Strategies

This method is implemented in PyTorch, building upon the 3D Gaussian Splatting framework gsplat [21]. The Gaussian primitives are initialized using approximately 500,000 point-cloud points predicted by VGGT-X. Camera parameters are also taken from VGGT-X, and the scene scale is defined as the maximum distance from any camera position to the mean camera location. Each Gaussian has learnable parameters: mean, log-scale, quaternion rotation, opacity logit, and degree-3 spherical harmonic (SH) color coefficients. Learning rates (LRs) for these parameters are summarized in Table I. An exponential LR scheduler decays the mean LR to 1% over 7,000 steps, with evaluations at 4,000 and 7,000 steps, and SH degree increases by 1 every 1,000 steps.

The MLP-based module use an LR of 1.5e-4. For the appearance-embedding module used in MLP-based pose optimization, the embedding dimension is 32, with an LR of 1e-2 for the embedding and 1e-3 for the head. λ_{SSIM} is the weight for the SSIM loss and is set to $\mathbf{0.2}$, and λ_{Depth}, the depth loss weight is set to $\mathbf{0.05}$. Global alignment of VGGT poses provides a consistent coordinate system, while per-image local pose refinement during evaluation runs for 100 iterations with an LR of 8e-4, decaying exponentially to 1%. This refinement uses a gradient-masked L1 objective.

Near and far planes are set relative to the scene scale (near = 0.05 × scene scale, far = 1e10). At each evaluation step, memory usage, Gaussian count, PSNR [22], SSIM [23], and LPIPS [24] are recorded. Also the Absolute Trajectory Error (ATE), which measures the difference between an estimated trajectory and its ground truth by first aligning the two trajectories to minimize the error, and then computing the error at each point and averaging it, often as a Root Mean Square Error (RMSE) along with the rotational errors are measured for the camera poses. This training setup jointly optimizes geometry, radiance (via staged SH activation), and camera parameters, with controlled scheduling, lightweight

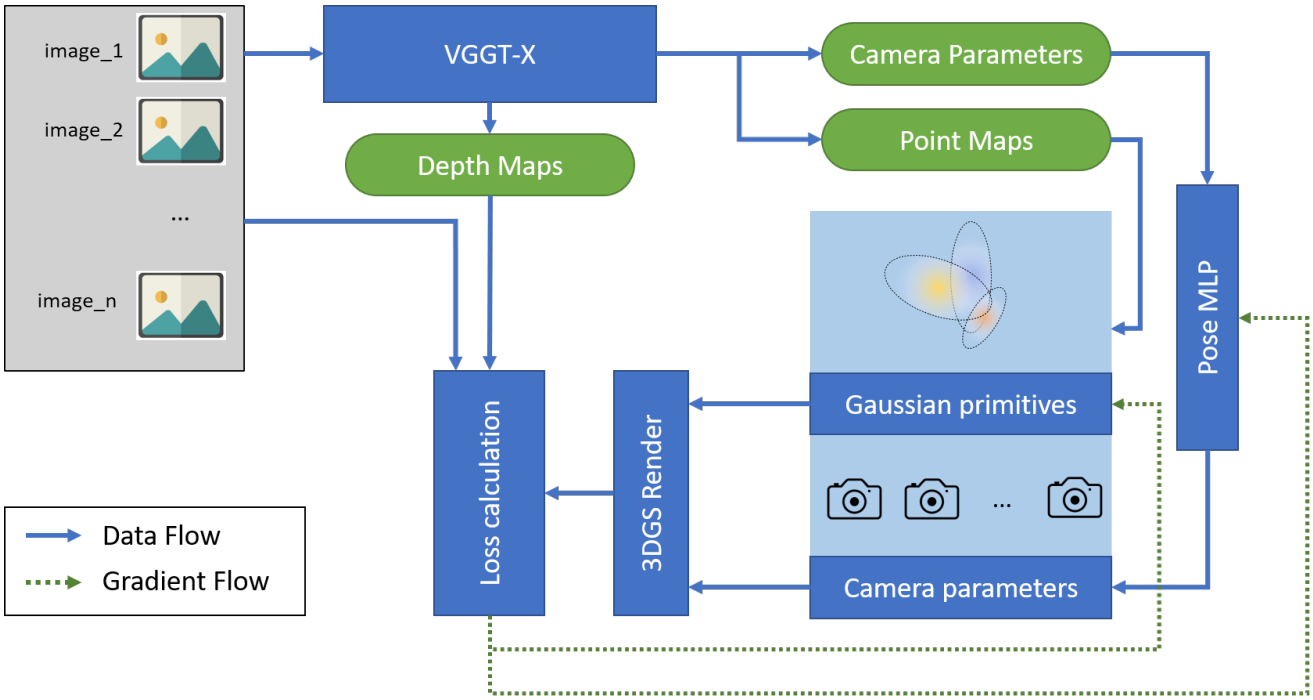

Fig. 2: FastSplatting Pipeline: Multi-view images are processed by VGGT-X to estimate camera poses and depth maps. These initial parameters seed the 3DGS model, which jointly optimizes Gaussian primitives and refines camera poses using a depth-guided loss function for accurate reconstruction from as few as 25 viewpoints.

per-image refinement, and optional correspondence-based constraints.

For calculating ATE and rotational error, COLMAP optimized camera parameters are used. The extrinsic from the COLMAP is transferred to the predicted coordinates using the Kabsch-Umeyama algorithm [25]–[27]. This is a method for finding the optimal transformation (translation, rotation, and scaling) to align two sets of corresponding points by minimizing the root-mean-square deviation (RMSD). This translation, rotation and scale is used to scale the COLMAP optimized extrinsics and is used as ground-truth.

TABLE I: Learning rates for Gaussian parameters

Parameter	Learning Rate
Mean	1.6e-4 times scene scale
Log-scales	5e-3
Quaternion rotation	1e-3
Opacity logit	5e-2
SH color coefficients (DC)	2.5e-3
SH color coefficients (higher orders)	1.25e-4

C. Results

All experiments are conducted on a machine with an NVIDIA RTX 4090 GPU and 64GB RAM. Table II summarizes baseline training performance across diverse indoor/outdoor scenes from the datasets mentioned above. All runs use 1M Gaussians, with 500,000 Gaussians initialized from VGGT-X. Memory differences stem from per-scene activation and view count. Higher PSNR/SSIM and lower LPIPS/ATE/RTE indicate better reconstruction and pose accuracy.

Table III compares novel view synthesis performance against 3DGS [1], ZeroGS [28], and 3RGS [9] across four scenes from Mip-NeRF 360. Our method achieves competitive PSNR, SSIM, and LPIPS scores, demonstrating high-fidelity reconstructions with better perceptual quality (LPIPS) while only training for 7,000 iterations, taking an average time of 180 seconds. Notably, it outperforms other techniques in LPIPS score, highlighting the effectiveness of our SfM-free pipeline and joint optimization strategy. Table IV compares the pose errors for the same scenes using the same methods.

In Figure 3, qualitative rendering results from FastSplatting on various scenes from the selected datasets are presented. The rendering are from the validation dataset. The images demonstrate high-fidelity reconstructions with accurate geometry and appearance details, showcasing the effectiveness of our SfM-free Gaussian splatting pipeline.

D. Ablation Studies

To quantify the contribution of each component, ablation studies were conducted based on: (1) 3DGS as an MCMC sampler, and (2) depth-guided loss, which is summarized in Table V. Evaluations are conducted on selected scenes from the datasets, reporting average metrics for both novel view synthesis and camera pose estimation. The results highlight substantial improvements from incorporating each component. Although the baseline 3DGS method already performs camera pose optimization during training, its performance remains limited without the proposed additions.

TABLE II: Baseline training metrics for training and validation set. Metrics with ↑ are higher-better; ↓ lower-better.

Scene	Time (s)	Mem (GB)	Train PSNR ↑	Train SSIM ↑	Train LPIPS ↓	Val PSNR ↑	Val SSIM ↑	Val LPIPS ↓	ATE ↓	RTE ↓
bicycle	154.18	9.21	22.26	0.7103	0.2003	24.22	0.7103	0.2003	0.0048	0.4078
drjohnson	143.16	9.92	23.63	0.5666	0.4842	19.69	0.5666	0.4842	0.0952	4.8973
stump	141.77	9.92	24.53	0.4966	0.2696	22.59	0.4966	0.2696	0.0117	0.3843
bonsai	173.54	9.58	26.89	0.7979	0.1266	25.99	0.7979	0.1266	0.0052	0.7599
garden	151.84	9.58	26.75	0.6122	0.1415	22.65	0.6122	0.1415	0.0016	0.3844
room	159.79	9.92	29.39	0.8921	0.0897	28.48	0.8921	0.0897	0.0073	0.6652
android	184.26	9.92	24.26	0.7977	0.1442	23.50	0.7977	0.1442	0.0093	0.5867
counter	167.78	9.58	26.80	0.8293	0.1097	26.16	0.8293	0.1097	0.0046	0.6045
crab1	178.50	9.92	25.65	0.8581	0.1725	23.80	0.8581	0.1725	0.0109	1.7140
playroom	140.36	9.92	25.95	0.7936	0.2020	25.04	0.7936	0.2020	0.0531	2.2704
kitchen	175.10	9.58	26.41	0.7485	0.1180	24.52	0.7485	0.1180	0.0052	0.4322
truck	166.74	9.92	23.57	0.8057	0.1275	22.28	0.8057	0.1275	0.0049	0.4924
crab2	249.77	16.48	24.97	0.8387	0.2045	24.12	0.8387	0.2045	0.0147	2.4636
yoda	281.29	16.72	26.15	0.8423	0.2197	25.33	0.8423	0.2197	0.0155	2.6221
statue	226.84	16.48	20.39	0.7374	0.2265	20.09	0.7374	0.2265	0.0077	0.2853
train	184.20	9.92	21.24	0.6839	0.2347	19.24	0.6839	0.2347	0.0057	0.5182
AVG	179.94	11.04	24.93	0.8228	0.1683	23.61	0.7507	0.1919	0.0161	1.2180

TABLE III: Quantitative comparison of novel view synthesis. (-) denotes unreported results for ZeroGS.

Scenes	3DGS			ZeroGS			3RGS			FastSplatting		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
garden	24.85	0.729	0.126	25.47	0.839	0.107	26.44	0.820	0.131	22.65	0.6122	0.1415
counter	27.57	0.862	0.209	26.87	0.873	0.124	28.80	0.897	0.157	26.16	0.8293	0.1097
bicycle	17.52	0.303	0.567	23.10	0.707	0.201	24.89	0.727	0.252	24.22	0.7103	0.2003
room	30.66	0.899	0.204	-	-	-	31.82	0.924	0.154	28.48	0.8921	0.0897

TABLE IV: Quantitative comparison of camera pose registration. (-) indicates unreported results.

Scenes	3DGS		ZeroGS		3RGS		FastSplatting	
	Rotation(°)↓	ATE(m)↓	Rotation(°)↓	ATE(m)↓	Rotation(°)↓	ATE(m)↓	Rotation(°)↓	ATE(m)↓
garden	0.19	0.003	0.03	0.002	0.03	0.002	0.3844	0.0016
counter	0.25	0.011	0.03	0.002	0.05	0.003	0.6045	0.0046
bicycle	1.07	0.034	0.04	0.005	0.09	0.013	0.4078	0.0048
room	0.27	0.016	-	-	0.13	0.012	0.6652	0.0073

TABLE V: Ablation study comparing training/validation reconstruction metrics and pose errors.

Scene	Train PSNR↑	Train SSIM↑	Train LPIPS↓	Val PSNR↑	Val SSIM↑	Val LPIPS↓	ATE↓	RTE↓
BASELINE	24.93	0.8228	0.1683	23.61	0.7507	0.1919	0.0161	1.2180
NO MCMC	21.58	0.7242	0.3030	20.99	0.6796	0.3063	0.0156	1.1560
NO DEPTH	25.87	0.8316	0.1679	21.58	0.6711	0.2584	0.0211	1.2549

V. Conclusion

In this study, FastSplatting is introduced, a novel pipeline that integrates 3D Feature Matching Transformers by means of 3DFMs VGGT-X with 3DGS to achieve ultra-fast and high-fidelity 3D scene reconstruction without relying on traditional SfM techniques. By leveraging VGGT-X for accurate camera pose and depth estimation, we effectively initialized the Gaussian splatting process, significantly reducing optimization time to around three minutes and improving convergence. The proposed MLP-based pose refinement and depth-guided loss further enhances reconstruction quality which achieves superior LPIPS score, particularly in low-viewpoint scenarios. Extensive experiments on diverse datasets demonstrated that FastSplatting outperforms existing methods in both novel view synthesis and camera pose accuracy, achieving competitive results with substantially reduced training times. Future work will explore extending this framework to dynamic scenes and integrating semantic understanding for enriched 3D reconstructions.

References

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, July 2023.

[2] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[4] J. Lee and S. Yoo, "Dense-sfm: Structure from motion with dense consistent matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6404–6414, 2025.

[5] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[6] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *CVPR*, 2024.

[7] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," 2024.

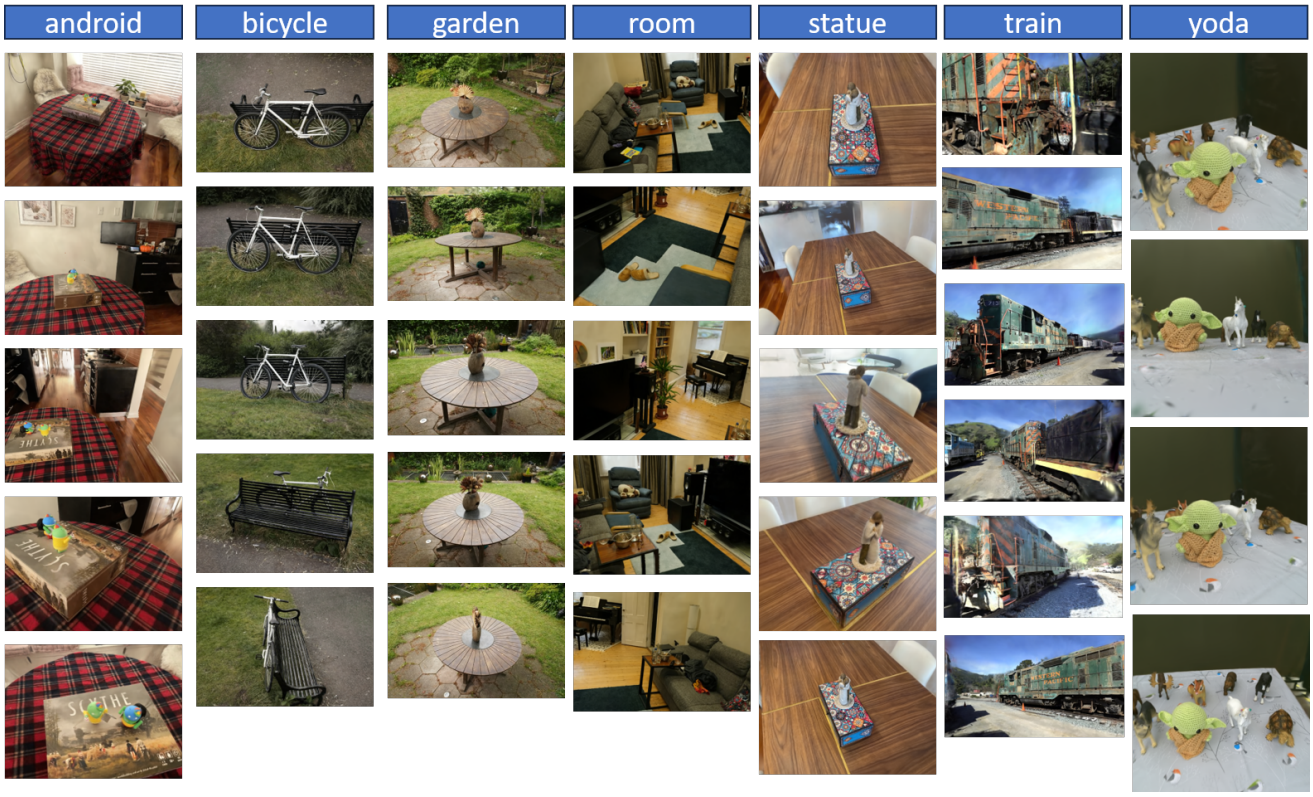

Fig. 3: Rendering results of FastSplatting on various scenes in the selected datasets.

[8] Y. Liu, C. Luo, Z. Tang, J. Peng, and Z. Zhang, "Vggt-x: When vggt meets dense novel view synthesis," 2025.

[9] Z. Huang, P. Wang, J. Zhang, Y. Liu, X. Li, and W. Wang, "3r-gs: Best practice in optimizing camera poses along with 3dgs," 2025.

[10] A. Dalal, D. Hagen, K. G. Robbersmyr, and K. M. Knausgård, "Gaussian splatting: 3d reconstruction and novel view synthesis: A review," *IEEE Access*, vol. 12, pp. 96797–96820, 2024.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[13] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[15] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," 2024.

[16] S. Kheradmand, D. Rebain, G. Sharma, W. Sun, J. Tseng, H. Isack, A. Kar, A. Tagliasacchi, and K. M. Yi, "3d gaussian splatting as markov chain monte carlo," 2025.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," *CVPR*, 2022.

[19] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[20] S. Sabour, S. Vora, D. Duckworth, I. Krasin, D. J. Fleet, and A. Tagliasacchi, "Robustnerf: Ignoring distractors with robust losses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20626–20636, June 2023.

[21] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, J. Ye, J. Hu, M. Tancik, and A. Kanazawa, "gsplat: An open-source library for gaussian splatting," *Journal of Machine Learning Research*, vol. 26, no. 34, pp. 1–17, 2025.

[22] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, IEEE, 2010.

[23] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[25] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Foundations of Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[26] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Foundations of Crystallography*, vol. 34, no. 5, pp. 827–828, 1978.

[27] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 4, pp. 376–380, 2002.

[28] Y. Chen, R. A. Potamias, E. Ververas, J. Song, J. Deng, and G. H. Lee, "Zerogs: Training 3d gaussian splatting from unposed images," in *arXiv*, 2024.