InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds

Zhiwen Fan*†[1,2], Wenyan Cong*[1], Kairun Wen*[3], Kevin Wang[1], Jian Zhang[3], Xinghao Ding[3], Danfei Xu[2,4], Boris Ivanovic[2], Marco Pavone[2,5], Georgios Pavlakos[1], Zhangyang Wang[1], Yue Wang[2,6]

* Equal contribution † Project leader

[1]University of Texas at Austin [2]Nvidia Research [3]Xiamen University
[4]Georgia Institute of Technology [5]Stanford University [6]University of Southern California

Figure 1. **Novel View Synthesis Comparisons (Sparse-View, Pose-Free)**. We introduce InstantSplat, an efficient framework for simultaneous pose estimation and novel view synthesis in unrestricted scenarios. This approach incorporates 3D priors derived from a dense stereo model, a streamlined framework capable of reconstructing 3D scenes and enabling view synthesis within one minute, for large-scale scenes. Moreover, our method markedly enhances both pose estimation accuracy and rendering quality.

Abstract

While novel view synthesis (NVS) has made substantial progress in 3D computer vision, it typically requires an initial estimation of camera intrinsics and extrinsics from dense viewpoints. This pre-processing is usually conducted via a Structure-from-Motion (SfM) pipeline, a procedure that can be slow and unreliable, particularly in sparse-view scenarios with insufficient matched features for accurate reconstruction. In this work, we integrate the strengths of point-based representations (e.g., 3D Gaussian Splatting, 3D-GS) with end-to-end dense stereo models (DUSt3R) to tackle the complex yet unresolved issues in NVS under unconstrained settings, which encompasses pose-free and *sparse view challenges. Our framework,* ***InstantSplat****, unifies dense stereo priors with 3D-GS to build 3D Gaussians of large-scale scenes from sparse-view & pose-free images in less than 1 minute. Specifically, InstantSplat comprises a Coarse Geometric Initialization (CGI) module that swiftly establishes a preliminary scene structure and camera parameters across all training views, utilizing globally-aligned 3D point maps derived from a pre-trained dense stereo pipeline. This is followed by the Fast 3D-Gaussian Optimization (F-3DGO) module, which jointly optimizes the 3D Gaussian attributes and the initialized poses with pose regularization. Experiments conducted on the large-scale outdoor Tanks & Temples datasets demonstrate that InstantSplat significantly improves SSIM (by 32%) while concurrently reducing Absolute Trajectory Error (ATE) by 80%. These establish InstantSplat as a viable solution for scenarios involving pose-free and sparse-view conditions. Project page:* https://instantsplat.github.io/.

1. Introduction

Novel-view synthesis (NVS) renders new images from unseen viewpoints of a scene based on a specific set of input images. Capturing viewpoints in a "casual" manner, especially with a limited number of shots (a.k.a. sparse-view),

is pivotal for scaling up 3D content creation, digital twin construction, and augmented reality applications.

Although recent advancements [10, 17, 20, 27] have shown notable progress in sparse-view synthesis (SVS). However, the sparse input data collected does not sufficiently cover the scene, preventing Structure from Motion (SfM) pipelines like COLMAP [19] from estimating accurate camera parameters. Previous research in SVS [10, 17] typically assumes precise camera poses even in sparse-view scenarios by leveraging all dense views for pre-computation, an assumption that is rarely valid. On the other hand, another line of research explores pose-free settings using techniques such as Neural Radiance Field (NeRF) or 3D Gaussian Splatting (3D-GS), exemplified by Nope-NeRF [5] and CF-3DGS [9]. These approaches presume dense data coverage (often from video sequences), an assumption that may not be viable in "casual" scenarios.

In this paper, we introduce a holistic solution to unconstrained sparse-view synthesis that obviates the need for accurately pre-computed camera intrinsics and extrinsics. We present InstantSplat, a framework that unifies the explicit 3D Gaussian representation with pose priors obtained from an end-to-end dense stereo model—DUSt3R [21]. DUSt3R facilitates the acquisition of initial and coarse scene geometry from predicted and globally aligned point maps of sparse views and enables efficient camera information and pose retrieval. Following this, fast 3D Gaussians adaptations is established to jointly optimize 3D Gaussian attributes and camera parameters. By merely adjusting Gaussian attributes—eschewing complex adaptive density control—the reconstruction process for large-scale scenes can be completed in under *one minute* on a modern GPU (Nvidia A100). Experiments on two large-scale outdoor datasets: Tanks & Temples [13] and MVImgNet [29], featuring sampled sparse views. Our evaluations, demonstrate InstantSplat remarkably surpasses previous pose-free method: the SSIM is boosted from 0.68 to 0.89, a 32% improvement, and the ATE is reduced from 0.055 to 0.011, while significantly accelerating the optimization (from $\sim$2 hours to approximately 1 minute) than Nope-NeRF [5].

2. Related Works

Novel view synthesis aims to render unseen views of an object or scene from a set of images [1, 15]. Neural Radiance Fields (NeRF)[16], employs Multilayer Perceptrons to represent 3D scenes. Subsequent works [2–4, 6, 12, 26] improve rendering quality or the efficiency. The recent 3D Gaussian Splatting(3D-GS) [12] uses anisotropic 3D Gaussians [32] to depict radiance fields which shows considerable success in rapidly reconstructing complex real-world scenes with high quality.

NeRFs and 3D-GS require over a hundred images as input and utilize preprocessing software, such as COLMAP [19], to compute camera intrinsics and extrinsics. The dense coverage of the capture images significantly limits practical applications. Works in reducing the view number requirements [8, 10, 17, 20, 24, 25, 27, 27, 28, 31] adopt either geometric regularizations or generative priors to preserve the rendering quality. However, these methods still require known ground-truth camera poses, a challenging prerequisite as the commonly used Structure-from-Motion (SfM) algorithms often fail with sparse inputs due to insufficient image correspondences. Several works [5, 7, 9, 11, 11, 14, 23] explore either utilize coarse-to-fine paradigm or utilizing monocular depth priors to optimize the camera parameters. But these works require dense multi-view coverage (e.g., video sequences).

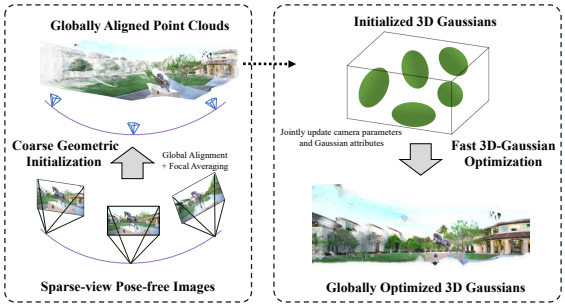


Figure 2. **Overall Framework of InstantSplat.** Starting with sparse, unposed images, the Coarse Geometric Initialization (left) rapidly predicts global aligned point clouds and initializes poses (20.6 seconds). Then the Fast 3D-Gaussian Optimization (right) leverages this initialization to conduct streamlined optimizations of 3D Gaussians and camera parameters (16.67 seconds).

3. Method

3.1. Coarse Geometric Initialization

Recovering Camera Intrinsics. We can have the 1:1 mapping from the pixel lattice to pointmap where we can build the mapping from 2D to the camera coordinate system. By utilizing Weiszfeld algorithm [18], we obtain the per-camera focal: $f^* = \arg\min_f \sum_{i=0}^{W}\sum_{j=0}^{H} \boldsymbol{O}^{i,j} \left\| (i',j') - f\frac{(\boldsymbol{P}^{i,j,0},\boldsymbol{P}^{i,j,1})}{\boldsymbol{P}^{i,j,2}} \right\|$, where $i' = i - \frac{W}{2}$ and $j' = j - \frac{H}{2}$ denote centered pixel indices. Assuming a single-camera setup akin to COLMAP's methodology, we average the focal length calculations to obtain a robust estimate: $\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f_i^*$. The resulting $\bar{f}$ represents the computed focal length that is utilized in subsequent processes.

Pair-wise to Globally Aligned Poses. Scaling from two-view to all views requires the alignment of the scale. We first construct a complete connectivity graph of all the N input views, and convert the initially predicted point map $\{(\boldsymbol{P}_i \in \mathbb{R}^{H\times W\times 3})\}_{i=1}^{N}$ to be globally aligned one $\{(\tilde{\boldsymbol{P}}_i \in$

$\mathbb{R}^{H\times W\times 3})\}_{i=1}^{N}$, by updating the point maps using transformation matrix, and a scale factor (refer more details in DUSt3R [21]). The post-processing yields a globally aligned point cloud but the optimized poses are still sub-optimal caused by the sub-optimal point maps predictions.

3.2. Fast 3D-Gaussian Optimization

3D Gaussian Initializations. To refine the initial camera poses and 3D geometry, we propose to utilize the optimization of the 3D Gaussians, and utilize photometric signals to jointly tune them into a more optimal solution. As we have per-pixel wise dense initialization, we propose to utilize the globally aligned point map as preliminary scene geometry, replacing the sparse SfM point set for 3D-GS initialization [12]. The ample primitives to encapsulate the scene's surfaces, minimizes the need for manual optimization rules (adaptive density control in 3DGS [12]), and thus requiring fewer steps. Specifically, we execute 1k iterations optimization on the initialized 3D Gaussians, omitting the densification, splitting, and opacity reset processes, thereby streamlining and simplifying the optimization procedure.

Jointly Optimizing Poses and Attributes. We mitigate the geometric and pose inaccuracy when transitioning from two-view to multi-view scenarios: we propose to simultaneous optimize the camera extrinsics and the 3D model using a sparse set of training views. Additionally, we introduce a constraint to ensure that the optimized poses do not deviate excessively from their initial positions:

$$\boldsymbol{S}^*, \boldsymbol{T}^* = \arg\min_{\boldsymbol{S},\boldsymbol{T}} \sum_{v\in N}\sum_{i=1}^{HW} \left\| \tilde{\boldsymbol{C}}_v^i(\boldsymbol{S},\boldsymbol{T}) - \boldsymbol{C}_v^i(\boldsymbol{S},\boldsymbol{T}) \right\| + \lambda \cdot \|\boldsymbol{T} - \boldsymbol{T0}\| .$$

the $\boldsymbol{S}$ represents the set of 3D Gaussians, $\boldsymbol{T}$ denotes the camera extrinsics for a specific view, $\boldsymbol{T0}$ signifies the initial extrinsics obtained from global alignment, $\boldsymbol{C}$ is the rendering function, and the term λ is introduced to strike a balance between photometric loss and the steps involved in camera optimization.

3.3. Aligning Camera Poses on Test Views

Test camera poses are unknown in the pose-free setting. Follow NeRFmm [23], we freeze the well-trained 3DGS model and optimize the camera poses for test views by matching the test images with the 3D model.

4. Experiments

4.1. Experimental Setup

Results are averaged on five scenes from the Tanks and Temples datasets [13], and extracted six outdoor scenes from the MVImgNet datasets. 12 uniformly sampled training and testing views are used. All methods are trained and tested under resolution of 512×288. Absolute Trajectory Error (ATE), the Relative Pose Error (RPE) are for pose accuracy, PSNR, Structural Similarity Index Measure (SSIM)[22], and the Learned Perceptual Image Patch Similarity (LPIPS)[30] are for rendering quality.

4.2. Experimental Results

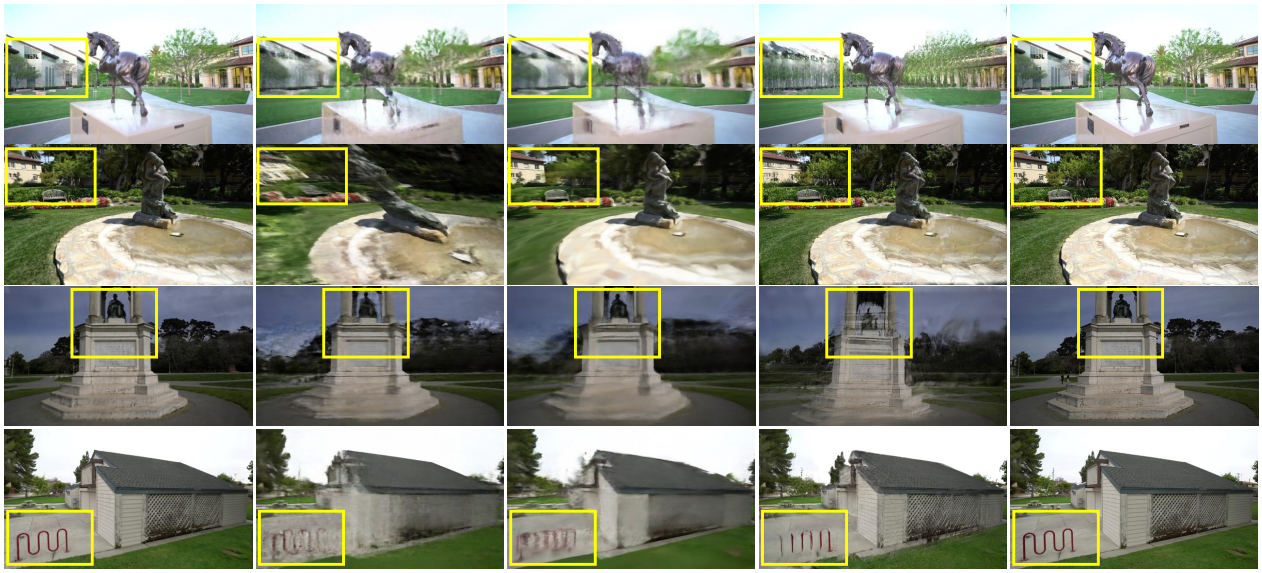
Quantitative and Qualitative Results Quantitative results of novel view synthesis and pose estimation on Tanks and Temples datasets, and MVImgNet, are summarized in Tab. 1. The pose metrics reveal inaccurate pose estimations attribute to sparse observations in these baseline methods. Specifically, Nope-NeRF [5] utilizes MLPs and achieves notable accuracy in rendering quality and pose estimation. However, it tends to produce overly blurred renderings (See Fig. 3) due to the heavy constraints from its geometric field and demonstrates slow training (∼2 hours) and inference speeds (∼30 seconds for one frame). CF-3DGS [9], employing Gaussian Splatting, delivers good rendering quality but is prone to artifacts when changing viewpoints, a consequence of the complex optimization process with erroneous pose predictions. Additionally, CF-3DGS requires more complex optimization, incorporating both local and global optimization stages, along with adaptive density control and opacity reset policies.

4.3. Ablation Studies

Effect of Averaging Focal Length and Joint Optimization. Experiments in Tab. 2 indicates that independent calculation of camera focal length results in a diversity of outcomes, adversely affecting rendering quality and pose estimation accuracy. The globally alignment algorithm does not yield pixel-wise accurate extrinsics, necessitate the joint optimization of camera parameters and Gaussian attributes.
Effect of View Number and Poses from COLMAP. We conduct experiments with varying numbers of training views. As illustrated in Tab 3, InstantSplat consistently outperforms CF-3DGS [9], another 3D Gaussian-based pose-free framework. Furthermore, COLMAP with vanilla 3D-GS also performs sub-optimal, caused by the problem that recover the entire 3D scene from sparse point cloud and training images.

5. Conclusion

We introduced InstantSplat, designed to reconstruct scene efficiently from sparse-view unposed images. Our approach leverages dense stereo priors for coarse scene initialization, offering preliminary estimates of the 3D geometry and camera parameters. To further refine these scene attributes and camera parameters, a rapid 3D Gaussian Optimization strategy that jointly optimizes the 3D Gaussian attributes and camera extrinsics. This results in an efficient pipeline capable of reconstructing the 3D scene from unposed images



Ours NeRFmm Nope-NeRF CF-3DGS Ground-Truth

Figure 3. **Visual Comparisons**. We conducted a comparative analysis between InstantSplat and various baseline methodologies. It was noted that InstantSplat adeptly preserves a majority of the scene details, avoiding the artifacts typically induced by inaccurately estimated camera poses, a common issue in CF-3DGS [9]. Moreover, our streamlined framework negates the necessity for strong regularization during training, unlike Nope-NeRF, thereby yielding sharper image details. Additionally, NeRFmm is prone to introducing artifacts during viewpoint transitions, attributable to imprecise joint optimization processes.

	Datasets	Ours			CF-3DGS			Nope-NeRF			NeRFmm		
		SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
Rendering Metrics	Tanks and Temples	0.89	28.94	0.12	0.60	18.29	0.34	0.64	22.48	0.44	0.51	18.28	0.53
	MVImgNet	0.79	25.19	0.19	0.38	16.63	0.45	0.45	19.14	0.55	0.27	13.88	0.66
		RPE_t ↓	RPE_r ↓	ATE↓	RPE_t ↓	RPE_r ↓	ATE↓	RPE_t ↓	RPE_r ↓	ATE↓	RPE_t ↓	RPE_r ↓	ATE↓
Pose Metrics	Tanks and Temples	0.472	0.110	0.011	5.797	2.175	0.070	10.279	5.303	0.055	15.026	4.701	0.125
	MVImgNet	0.317	0.279	0.004	10.119	9.083	0.106	12.501	13.700	0.142	15.014	11.281	0.132

Table 1. **Quantitative Evaluations.** Our method renders significantly clearer details (by LPIPS) compared to other baseline methods, devoid of artifacts typically associated with noisy pose estimation (e.g., CF-3DGS [9], NeRFmm [23]). Furthermore, Nope-NeRF's regularization approach during training, which involves multiple constraints, restricts the MLPs' ability to accurately reconstruct scene details. Results are on both 12 training and testing views.

Scenes	No Averaging Focal			
	PSNR↑	SSIM↑	RPE_t ↓	RPE_r ↓
No Focal Avg.	27.18	0.8552	1.053	0.135
No Joint Opt.	26.82	0.8547	0.677	0.173
Full Model	28.58	0.8900	0.472	0.110

Table 2. Ablation Study on the Impact of Averaging Focal Length and Jointly Optimization.

Scenes	InstantSplat				CF-3DGS				COLMAP+3DGS			
	3-view	5-view	9-view	12-view	3-view	5-view	9-view	12-view	3-view	5-view	9-view	12-view
Barn	0.778	0.821	0.886	0.884	0.351	0.405	0.516	0.544	0.582	0.687	0.790	0.796
Family	0.848	0.896	0.907	0.908	0.325	0.403	0.579	0.580	0.561	0.663	0.693	0.739
Francis	0.815	0.883	0.902	0.902	0.345	0.443	0.572	0.574	0.509	0.640	0.693	0.689
Horse	0.838	0.889	0.904	0.904	0.441	0.477	0.573	0.586	0.496	0.581	0.674	0.686
Ignatius	0.710	0.795	0.824	0.830	0.195	0.291	0.634	0.724	0.617	0.691	0.786	0.823
Avg.	0.798	0.857	0.885	0.886	0.331	0.404	0.575	0.602	0.553	0.652	0.727	0.747

Table 3. **Ablation Study on the view number.** We compare between InstantSplat, CF-3DGS, and COLMAP with vanilla 3D-GS. SSIM is used to report the rendering quality. InstantSplat consistently outperform other baselines.

in under one minute. Significantly, our method demonstrates superior rendering quality and pose estimation accuracy compared to existing methodologies, underscoring its effectiveness in handling sparse-view data.

Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Shai Avidan and Amnon Shashua. Novel view synthesis in tensor space. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1034–1040. IEEE, 1997. 2

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5460–5469, 2022. 2

[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.

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 2

[5] 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, 3

[6] Zhang Chen, Zhong Li, Liangchen Song, Lele Chen, Jingyi Yu, Junsong Yuan, and Yi Xu. Neurbf: A neural fields representation with adaptive radial basis functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4182–4194, 2023. 2

[7] Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhransu 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

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021. 2

[9] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. *arXiv preprint arXiv:2312.07504*, 2023. 2, 3, 4

[10] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[11] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021. 2

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3

[13] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 2, 3

[14] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16539–16548, 2023. 2

[15] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[16] 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

[17] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. *arXiv preprint arXiv:2112.00724*, 2021. 2

[18] F Plastria. The weiszfeld algorithm: proof, amendments and extensions, ha eiselt and v. marianov (eds.) foundations of location analysis, international series in operations research and management science, 2011. 2

[19] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[20] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *arXiv preprint arXiv:2303.16196*, 2023. 2

[21] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2312.14132*, 2023. 2, 3

[22] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[23] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2, 3, 4

[24] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. *arXiv preprint arXiv:2312.02981*, 2023. 2

[25] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2

[26] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5438–5448, 2022. 2

[27] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8254–8263, 2023. 2

[28] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[29] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 2

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[31] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00451*, 2023. 2

[32] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS'01.*, pages 29–538. IEEE, 2001. 2