Intern-GS: Vision Model Guided Sparse-View 3D Gaussian Splatting

XIANGYU SUN

M.Phil

THE UNIVERSITY OF
SYDNEY

Supervisor: Dr. Tongliang Liu
Associate Supervisor: Dr. Baosheng Yu

A thesis submitted in fulfilment of
the requirements for the degree of
Master of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

1 July 2025

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Student Name: Xiangyu Sun **Date:** 1 July 2025

Student Signature:

Authorship Attribution Statement

I declare that the thesis is my own original work and has not been previously published, in whole or in part, for a degree or any other qualification at this or any other institution. I confirm that this thesis does not contain any material previously published or written by myself or others, except where due reference is made in the text of the thesis. All sources of information and assistance have been specifically acknowledged.

Student Name: Xiangyu Sun **Date:** 1 July 2025

Student Signature:

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name: Tongliang Liu **Date:** 1 July 2025

Supervisor Signature:

Statement of Use of Gen AI

During the preparation of this thesis, Microsoft Copilot Chat was used for text enhancement purposes, including grammatical corrections. For example, I modified the sentence "These limitations typically incomplete information, which can lead to suboptimal reconstructions using existing methodologies" to "These limitations result in incomplete information, leading to suboptimal reconstructions using existing methodologies" for grammatical correctness. I have thoroughly reviewed all text modified by generative AI to correct any errors, inaccuracies, or biases, and made adjustments as needed. I take full responsibility for the submitted thesis, ensure the work is my own, and have used generative AI within acceptable parameters.

Student Name: Xiangyu Sun **Date:** 1 July 2025

Student Signature:

Abstract

Sparse-view scene reconstruction often faces significant challenges due to the constraints imposed by limited observational data. These limitations result in incomplete information, leading to suboptimal reconstructions using existing methodologies. To address this, we present Intern-GS, a novel approach that effectively leverages rich prior knowledge from vision models to enhance the process of sparse-view Gaussian splatting, thereby enabling high-quality scene reconstruction. Specifically, Intern-GS utilizes vision foundation models to guide both the initialization and the optimization process of 3D Gaussian splatting, effectively addressing the limitations of sparse inputs. In the initialization process, our method employs DUSt3R to generate a dense Gaussian point cloud. This approach significantly alleviates the limitations encountered by traditional structure-from-motion (SfM) methods, which often struggle under sparse-view constraints. During the optimization process, vision foundation models predict depth and appearance for unobserved views, refining the 3D Gaussians to compensate for missing information in unseen regions. We have tested Intern-GS across a wide range of datasets, encompassing both forward-facing and large-scale scenes. Our experiments demonstrate that Intern-GS consistently achieves state-of-the-art rendering quality on the LLFF dataset, the DTU dataset, and the Tanks and Temples dataset.

Acknowledgements

Time flies, my graduate careeris about to come to an end. At this moment, my heart is filled with deep humility and endless gratitude. This academic journey is undoubtedly a process of growth, and my gains and progress are inseparable from the support and help of many people. Here, I sincerely thank all the professors, workmates, family and friends who have given me guidance, companionship and encouragement in my study and research.

First of all, I would like to express my deepest respect and heartfelt gratitude to Professor Tongliang Liu. During my Mphil studies, Professor Liu became a light on my academic path with his rigorous academic attitude, profound knowledge and keen academic insight. His guidance played a vital role in my academic growth. When I encountered research difficulties or fell into confusion, he always gave me clear ideas and clear directions, patiently answered my questions, and provided valuable advice and guidance. Without his selfless help and meticulous guidance, I can hardly imagine that I could successfully complete this thesis.

At the same time, I am also deeply grateful to my co-supervisor Dr. Runnan Chen. Under his careful guidance, I not only gained academic growth, but also gained precious friendships. Throughout the research process, Dr. Chen always gave me meticulous care and profound guidance. His academic advice has benefited me a lot, and his help in experimental design, data analysis, and paper writing is indispensable. His patience and rigorous attitude helped me to overcome difficulties and find better solutions.

In addition, I would like to express special thanks to my family and girlfriend. It is their unwavering support and encouragement that keep me moving forward. I am sincerely grateful to my parents, who have given me unlimited strength with their selfless dedication and strong support. Their understanding and encouragement are the source of my continuous pursuit of excellence. At the same time, I also cherish the company and care of my girlfriend. She always accompanies me silently and never asks me for anything. In difficult times, their support keeps me optimistic and firm.

The completion of this thesis is not only the result of my personal efforts, but also the hard work and dedication of all those who have supported and helped me. Here, I once again express my most sincere gratitude to all those who have helped and encouraged me. The road ahead is still long. With a grateful heart, I will continue to pursue excellence and work hard to live up to all those who trust and support me.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

3D reconstruction [57] is a transformative and multidisciplinary field dedicated to the creation of highly detailed three-dimensional representations of objects, scenes, or environments by utilizing 2D images or other supplementary data sources. This intricate process bridges the gap between static, flat imagery and dynamic, immersive virtual models, enabling the generation of lifelike digital reconstructions that closely mimic real-world entities [12]. The significance of 3D reconstruction lies in its vast potential to revolutionize a multitude of industries and applications. For instance, in virtual reality (VR) and augmented reality (AR), it facilitates the creation of immersive environments that enhance user experiences. In film production, it enables the generation of realistic visual effects and digital doubles. In urban planning, it provides accurate 3D models of cities for simulation and development. In the field of autonomous driving, 3D reconstruction technology supports the construction of high-precision maps and the development of environmental perception systems. In addition, in game design, it makes it possible to build virtual worlds with rich details. Given its wide influence and versatility, 3D reconstruction has become a vital and rapidly developing research direction in the field of computer vision, constantly pushing the technical boundaries of digital representation and simulation.

From the 3D reconstruction of a scene, we can render some novel 2D images in unseen views, which is called Novel View Synthesis (NVS) [61, 3]. NVS focuses on generating images from unseen perspectives using a series of images captured from specific viewpoints. Traditional NVS methods involve creating 3D models from radar scan data or multi-view image data using specialized software to generate new perspectives. While this 3D modeling approach can produce highly realistic 3D content, it requires professional expertise and real-time user

interaction, which is time-consuming [50]. Additionally, imperfect camera parameters may introduce noise into the results. Recent advances in NVS, particularly those which use Neural Radiation Fields (NeRF) [29] and 3D Gaussian splatting [19], have significantly advanced the capabilities of NVS technologies [14, 25, 62]. NeRF utilizes an implicit neural network to model the density at various points in space and employs volume rendering techniques to reconstruct the 3D scene's geometry and appearance. It has significantly improved the quality of novel view synthesis results, but still faces challenges related to slow training and rendering speeds. To address the issue of speed, 3D Gaussian Splatting employs a set of Gaussian ellipsoids for rasterization to approximate the appearance of a 3D scene. This method achieves real-time rendering while maintaining comparable quality in novel view synthesis.

However, in many real-world applications, obtaining dense views is not feasible, and the available views are usually sparse, covering only a limited number of perspectives [16, 4, 53, 9]. This sparse-view scenario poses substantial challenges as the large number of unobserved viewpoints results in significant information gaps, which critically impact the completeness and quality of reconstructions. This is because these sophisticated methods mentioned above generally require dense input views [9] and precise camera poses [16, 56], and often begin with sparse point clouds derived from advanced Structure-from-Motion (SfM) techniques [44, 38]. Sparse inputs often lack sufficient overlap, which hampers the ability of SfM to estimate camera parameters accurately, causing it to struggle under sparse-view conditions. Particularly in areas of poor texture and smooth surfaces, SfM frequently fails to accurately match features across multiple images. This often results in rendered scenes that are plagued by artifacts and inconsistencies [30, 8, 56]. The fundamental issue here stems from a lack of sufficient prior information. A straightforward approach to address this is to provide the model with more accurate and robust prior information. In this context, vision foundation models [34, 32, 24], which are pretrained on extensive and diverse datasets, present a promising avenue by providing comprehensive visual priors that significantly help bridge information gaps in sparse-view NVS.

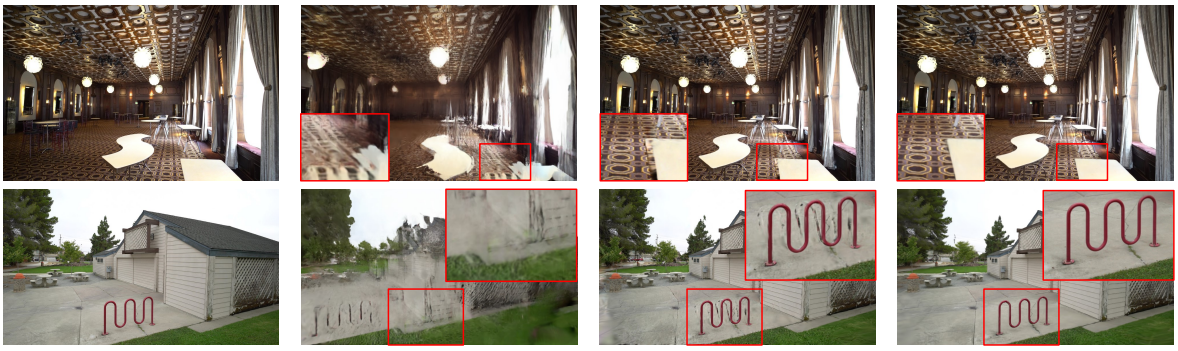

FIGURE 1.1. Comparison of the state-of-art SparseNeRF [45], original 3D Gaussian [19] in 3 training views. We introduce Intern-GS, an efficient framework for 3D Gaussian to reconstruct scene in sparse views. Our work leverage multi-view stereo prior to densely initial 3D Gaussian, supervised using a combination of various forms of regularization. Our work significantly improved the quality of rendering.

To effectively address the challenges of sparse view environments, we introduce Intern-GS, a pioneering approach that leverages visual grounding models to guide the initialization and optimization of Gaussian representations. Our innovative strategy aims to address the inherent limitations of sparse view scenes by leveraging the rich priors embedded in the visual grounding models. These priors enable our method to bridge the information gap caused by the lack of input views, facilitate the generation of dense Gaussian initializations, and significantly enhance the refinement of depth and appearance from unobserved viewpoints. Specifically, our method starts with DUSt3R [47], a state-of-the-art multi-view stereo model, to create an initialization dense point cloud. Unlike conventional SfM techniques that typically struggle in textureless regions, DUSt3R leverages strong multi-view stereo priors to generate comprehensive and accurate point cloud initialization. This initial dense point cloud lays a solid foundation for further optimization, which is crucial in scenes with limited geometric or photometric data. During the optimization process, Intern-GS employs advanced visual models to improve the appearance and depth of unobserved viewpoints. We leverage a pre-trained diffusion model [46] to enhance appearance by generating realistic textures for areas without direct observation. Meanwhile, a state-of-the-art deep depth estimation model predicts accurate depth values, providing key geometric constraints that guide our Gaussian optimization to achieve accurate 3D geometry for all view angles. By fusing these strategies (enhancing appearance with a diffusion model and improving depth with deep learning),

Intern-GS achieves superior detail and accuracy, setting a new standard for 3D reconstruction and novel view synthesis in challenging environments. Intern-GS has been rigorously tested on a variety of sparse view datasets and performs well in both forward and large-scale scenes. It significantly outperforms existing methods on challenging benchmarks such as LLFF [28], DTU [1], and Tanks and Temples [21].

This paper is well-structured and provides an in-depth exploration of our research on new view synthesis (NVS) under sparse view conditions. The chapters are laid out as follows: Chapter 2 provides a comprehensive literature review. It first introduces 3D scene representation methods for new view synthesis, and then outlines popular 3D scene reconstruction and NVS algorithms in recent years. Special emphasis is placed on 3D scene reconstruction and NVS techniques under sparse view conditions. Chapter 3 provides a detailed presentation of our proposed method. This includes the problem definition, the specifics of dense and non-redundant point cloud initialization during the initialization process, and its implementation. Additionally, it covers the optimization process, which involves multi-view stereo color and geometric regularization, as well as monocular depth regularization. This section also elaborates on the integration details of the diffusion model and the deployment of the monocular depth estimation model. Chapter 4 focuses on the experimental aspects of our research. It outlines the experimental setup, the datasets used, and the comparative methods. We conducted experiments not only on forward-facing datasets but also on large-scale real-world scenes, followed by extensive ablation studies. Chapter 5 draws conclusions based on our research findings, summarizing our study and highlighting the key contributions and their implications for the broader academic and research community.

Our contributions are substantial and can be highlighted as follows:

- We introduce Intern-GS, an innovative method for dense Gaussian initialization that utilizes multi-view stereo priors from vision foundation models, achieving unprecedented geometric consistency and outperforming traditional methods, especially in areas with low texture.
- We have developed a unique regularization mechanism that employs a pre-trained diffusion model for sophisticated appearance refinement and a deep depth estimation model to

enforce depth constraints, optimizing the Gaussian representations to ensure uniform color and depth across unobserved viewpoints.

- Our extensive experiments confirm Intern-GS as a cutting-edge solution for sparse-view NVS, excelling on demanding datasets such as LLFF, DTU, and Tanks and Temples, and setting a new gold standard for the field of sparse-view novel view synthesis.

CHAPTER 2

Literature Review

2.1 3D Representations

The objective of novel view synthesis (NVS) [61, 3] is to create images of scenes or objects from viewpoints that have not been directly observed, using data collected from specific known viewpoints. Recent advancements in this field have been particularly exciting, with Neural Radiance Fields (NeRF) [29] emerging as a significant breakthrough.

NeRF employs a multilayer perceptron (MLP) [36] to map spatial coordinates and viewing angles to corresponding colors and densities, enabling image rendering through volumetric techniques. Since its inception, efforts to refine NeRF have concentrated on enhancing the quality [29, 4, 58, 10] and efficiency [55, 15, 33, 7, 23] of the renderings, as well as improving 3D generation capabilities [42, 6, 31, 27, 39] and refining pose estimation techniques [41, 35, 64]. However, achieving real-time rendering performance continues to be a significant hurdle, as NeRF inherently requires extensive computational resources and processing time. In response to these limitations, the development of 3D Gaussian techniques [19] has gained traction. This approach involves initializing a series of anisotropic 3D Gaussians to model the entire radiance field comprehensively, followed by rendering the scene using differentiable splatting methods. This technique has proven highly effective in rapidly and accurately reconstructing complex real-world scenes, delivering robust performance under multi-view input scenarios [42]. Despite these advancements, numerous challenges persist, especially when dealing with sparse view inputs. Key issues include determining the appropriate prior constraints to apply and devising effective strategies for their implementation. These challenges necessitate a nuanced understanding of both the theoretical and practical aspects

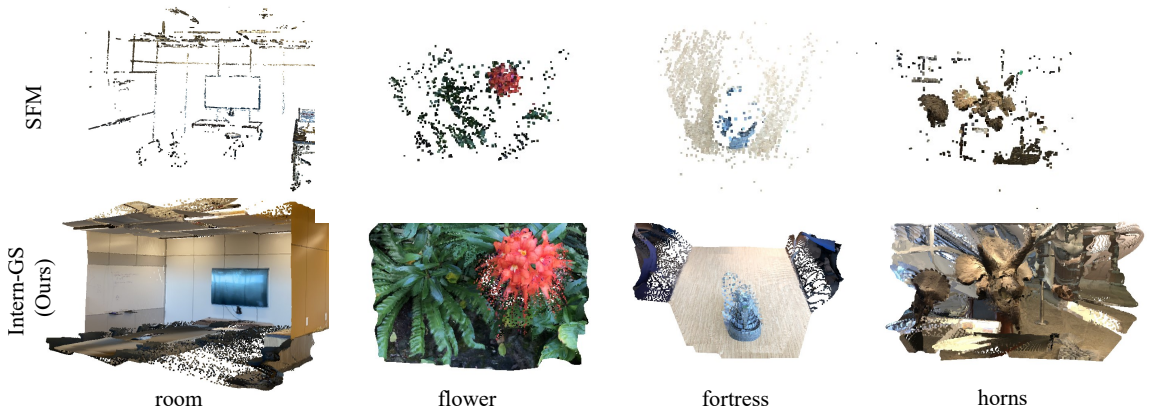

FIGURE 2.1. Comparison of Points Initialization of Original 3D Gaussian [19] and Our Method in 4 scenes using 3 train views. The first row's results are derived from Structure from Motion (SfM) [44] used by the original 3D Gaussian and most NeRF-based methods. In contrast, the second row shows the results of our initialization method. Obviously, our method outperforms the SfM method in texture-poor areas.

of 3D reconstruction and the innovative application of Gaussian models to overcome the inherent difficulties presented by sparse data inputs.

However, both of these popular methods exhibit significant flaws: they rely on dense input and accurate estimation of camera parameters. Camera poses are typically obtained through a separately optimized Structure from Motion (SfM) [44] process. SfM is a complex computational process that reconstructs three-dimensional structures by analyzing multiple 2D images taken from different viewpoints. This technique depends on identifying common feature points across multiple images and then using the geometric relationships between these points to infer the positions and orientations of the cameras, thereby localizing points in three-dimensional space. As the number of input views decreases, the accuracy of the estimated camera poses also diminishes, leading to a degradation in the quality of the SfM process, which can, in more severe cases, lead to its collapse.

2.2 Novel View Synthsis In Sparse-View

To combat the issue of reduced accuracy that arises from the diminished matches in corresponding points in Structure from Motion (SfM) as the number of input views decreases,

numerous studies have employed strategies that incorporate additional prior information or have designed specific regularization terms to enhance the performance of NeRF [20, 10, 60, 51, 20] and 3DGS [63, 52]. For example, DietNeRF [16] utilizes semantic priors derived from the expansive semantic model CLIP to introduce sophisticated semantic constraints in high-dimensional spaces, ensuring multi-view consistency in the rendered images of previously unseen views. Mip-NeRF 360 [4] extends NeRF's capabilities in processing wide-angle and panoramic views by integrating mipmapping techniques, which optimize angular continuity during the rendering process. This enhancement not only significantly boosts anti-aliasing performance but also preserves finer details more effectively. RegNeRF [30] focuses on improving the training stability of NeRF and mitigating the overfitting issues prevalent in models dealing with sparse view inputs by implementing a strategic depth regularization approach. SparseNeRF [45] addresses the memory and computational demands of NeRF when processing sparse view data by employing learnable sparse radiance probes, which streamline the model's efficiency. For Gaussian models, there are also some works under sparse-view conditions. FSGS [63] uses monocular depth priors as conditions in sparse-view environments, while sparseGS [52] employs diffusion models as its working mechanism.

However, while these advancements have made significant strides in refining NeRF's functionality, the potential of leveraging prior information from foundational visual models during the initialization phase to augment the matching of corresponding points remains largely unexplored.

2.3 Pose-Free Sparse-View NVS

The inaccuracies in initialization mainly stem from two sources: one is the insufficient density of point clouds initialized by SfM in sparse-view conditions, and the other is the inaccuracy of pose estimation during the SfM process due to a lack of sufficient information. In the second area, some research has already been conducted. Common approaches typically assume that the camera parameters are ground truth; however, SfM clearly struggles to perform well under sparse-view conditions. Therefore, these methods focus on pose-less optimization, aiming to

produce calibrated images as output. NeRF– [49] was a trailblazer in the joint optimization of camera intrinsics, extrinsics, and NeRF parameters. Following this, BARF [22] refined the approach with a coarse-to-fine strategy that enhanced both pose refinement and scene reconstruction. GARF [17] showcased the effectiveness of Gaussian-MLPs in optimizing both pose and scene parameters. Sparf [43] simultaneously optimizes NeRF's parameters and rectifies noisy poses, crafting consistency constraints across different viewpoints to mitigate the negative impacts of inaccurate poses on the model's optimization. More recent initiatives like Nope-NeRF [5] and CF-3DGS [13] have incorporated depth information to enhance the optimization processes for NeRF and 3DGS, thereby significantly improving the accuracy and robustness of neural scene representations. But they overlook the critical importance of point clouds as an initialization base. How to construct a setup where both the pose and the point cloud information are precise and richly detailed presents a worthwhile problem for further investigation.

CHAPTER 3

Methodology

3.1 Problem Definition

Our work aims to enhance the rendering of new views in sparse input settings by addressing the limitations of current Structure from Motion (SfM) based 3D Gaussian methods, which produce sparse point clouds in regions with sparse textures. This causes the model to be unable to learn accurate geometry in these regions, resulting in rendering artifacts such as floaters and ghosting. To alleviate this issue, we use multi-view stereo priors from a visual grounding model to create a denser point cloud during initialization, especially in regions with sparse textures. To reduce computational overhead and maintain rendering quality, we develop a novel approach to remove redundant points. Furthermore, we integrate depth regularization from multi-view stereo and photometric regularization from a diffusion model to ensure that the model accurately captures geometry and appearance from unseen viewpoints.

3.2 Preliminary of 3D Gaussian Splatting

3.2.1 Representation

The process of 3D scene representation typically begins with the extraction of a 3D point cloud from a set of input images, which is commonly achieved through SfM techniques [37]. This foundational step provides the initial spatial structure of the scene, which is subsequently refined and enhanced. In the context of 3D Gaussian Splatting, each point in the extracted point cloud is initialized as a corresponding 3D Gaussian distribution, denoted

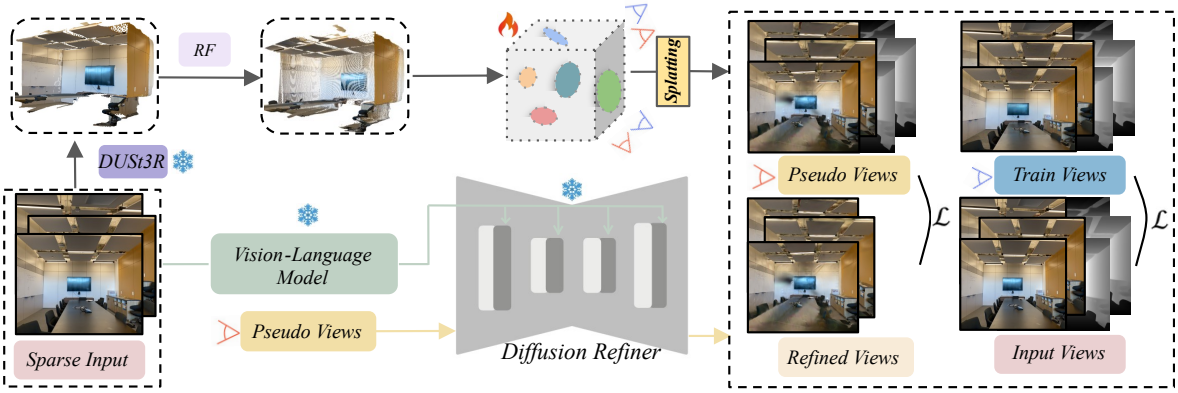

FIGURE 3.1. In our framework, we use DUSt3R [47] to predict point maps from multi-pair images. This technique recovers point maps at a consistent scale, but failed to represent scene because of redundancy in points. To handle overlapping regions in the point maps, we designed a Redundancy-Free (RF) algorithm that only initializes areas which have not been well defined for all views. This ensures a densely initialized and non-redundant 3DGS with accurate position. To further alleviate blurriness under the new perspective, we have designed a novel regularization method that jointly constrains the depth and color information of pseudo-views. The color supervision is derived from the additional diffusion refine model we employ, while the depth supervision comes from multi-view stereo.

as $G(x)$. Each individual Gaussian distribution $G_i(x)$ is characterized by a set of optical and geometric properties that collectively define its appearance and spatial configuration. The optical properties of each Gaussian include its opacity α_i and color c_i. The color c_i is represented by a spherical harmonic function of dimension l, which captures the directional dependence of the color and is expressed as: $\{c_i \in \mathbb{R}^3 \mid i = 1, 2, \ldots, l^2\}$ [19].

In 3D space, the position and shape of each Gaussian distribution are determined by its mean μ_i, which represents the center position of the Gaussian, and its covariance matrix Σ_i, which describes the spread and orientation of the distribution. The covariance matrix Σ_i is derived from two key parameters: the scale s_i and the rotation r_i. Specifically, the covariance matrix is computed as: $\Sigma = RSS^TR^T$, where R is a rotation matrix represented by quaternions, and S is a diagonal scaling matrix.

Mathematically, the i-th Gaussian distribution $G_i(x)$ is represented by the following equation:

$$G_i(x) = e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}, \tag{3.1}$$

where x denotes a point in 3D space, μ_i is the mean vector, and Σ_i is the covariance matrix.

The optimizable parameters for each Gaussian $G_i(x)$ include its opacity α_i, color c_i, mean position μ_i, rotation r_i, and scale s_i. Overall, the optimizable parameters of the i-th Gaussian $G_i(x)$ are $\{\alpha_i, c_i, \mu_i, r_i, s_i\}$. These parameters are iteratively refined during the optimization process to achieve an accurate and visually plausible representation of the 3D scene.

3.2.2 Rendering and Optimization Process

To compute the color of each pixel in the rendered 2D image, 3D Gaussian Spl-natting employs a rasterization process that aggregates contributions from N overlapping Gaussians influencing the pixel. The synthesized color C_p of pixel p is given by:

$$C_p = \sum_{i \in N} c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \tag{3.2}$$

where c_i represents the color contribution of the i-th Gaussian, computed from its spherical harmonics (SH) coefficients, and α'_i denotes the effective opacity of the i-th Gaussian in the 2D-pixel coordinate system. The effective opacity α' is derived from the projected covariance matrix Σ' in the 2D plane and the original opacity α of the 3D Gaussian. The projection of the 3D Gaussian into the 2D pixel coordinate system is achieved through a transformation:

$$\Sigma' = JW\Sigma W^T J^T, \tag{3.3}$$

where J is the approximate Jacobian matrix of the projection transformation, and W represents the rotational component of the camera pose.

A similar approach is used to render depth values for each pixel. The depth D_p of pixel p is computed as:

$$D_p = \sum_{i \in N} d_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \tag{3.4}$$

where d_i represents the depth value associated with the i-th Gaussian, calculated as the Euclidean distance between the Gaussian's mean position μ_i and the camera center o:

$$d_i = \|\mu_i - o\|_2. \tag{3.5}$$

3D Gaussian Splatting optimizes its model parameters through color constraints. During the optimization process, the algorithm dynamically adapts the Gaussian distributions by cloning or splitting them based on their gradient magnitudes and scales. Specifically, it clones Gaussians whose gradients exceed a predefined threshold but whose scales remain below a certain limit. At the same time, it splits Gaussians with gradients and scales, both exceeding their respective thresholds. This adaptive optimization strategy ensures a balance between detail preservation and computational efficiency. In our work, we retain the original optimization methodology and color constraints.

3.3 Multi-View Stereo Guided Dense Initialization

Although SfM methods are commonly used for initialization, NeRF and 3D Gaussian Splatting have significantly different reliance on the quality of initialization due to differences in the underlying algorithms. In particular, 3D Gaussian Splatting has weak color extrapolation capabilities in occluded areas and is therefore highly dependent on dense and precisely positioned point clouds. However, existing initialization techniques usually generate sparse correspondences that lack sufficient geometric details and make it difficult to fully exploit color priors. In addition, SfM-based methods are computationally intensive, limiting their feasibility in real-time applications. This phenomenon highlights the urgent need to design more efficient and robust initialization strategies for 3D Gaussian Splatting, especially for applications in real-time high-fidelity scene reconstruction.

3.3.1 Dense Stereo Point Cloud

In order to overcome the inherent limitations of sparse viewpoint cloud initialization in the SfM process, using dense point clouds is a direct and effective solution. Advances in deep learning technology have accelerated the development of Multi-View Stereo (MVS) frameworks, enabling them to integrate these innovative methods. Among them, DUSt3R [47] is a typical representative of this progress, which can instantly generate 3D point cloud maps and confidence maps based on only two input images. The framework establishes an

accurate one-to-one correspondence between 2D image pixels and points in 3D scene space, and its mathematical description is as follows:

$$I_p \leftrightarrow X_p, \tag{3.6}$$

where I denotes a pixel in the pixel coordinate system, X denotes a corresponding 3D point in the camera coordinate system, and p indexes the views. DUSt3R's training objective is specifically geared towards regressing between the original pixel maps and the predicted point maps derived from the two input views. The regression loss for a given view $p \in 1, 2$ is defined as the Euclidean distance:

$$\ell_{\text{regr}}(p) = \left| \frac{1}{z} X_{p,1} - \frac{1}{\hat{z}} \hat{X}_{p,1} \right|, \tag{3.7}$$

where $\hat{X}$ represents the ground-truth points, and X represents the predictions. To address the scale ambiguity between the predictions and the ground-truth data, DUSt3R applies scaling factors $z = \text{norm}(X^{1,1}, X^{2,1})$ and $\hat{z} = \text{norm}(\hat{X}^{1,1}, \hat{X}^{2,1})$ to normalize the predicted and ground-truth point maps respectively.

In addition, DUSt3R improves its prediction accuracy by learning to generate confidence maps, where each score quantifies the network's confidence in the accuracy of the corresponding pixel. This approach is particularly effective in reducing the problem of blurry 3D points such as sky areas or transparent objects. To optimize this ability, a weighted confidence regression loss is introduced on all pixels during training, with the following optimization objectives:

$$L_{\text{conf}} = \sum_{p \in 1,2} \sum_{i} C_i^{p,1} \ell_{\text{regr}}(p, i) - \alpha \log C_i^{p,1}, \tag{3.8}$$

where $C_i^{p,1}$ denotes the confidence score of pixel i, and α is a regularization hyper-parameter defined within the DUSt3R model. This structured approach ensures that both geometric accuracy and model confidence are finely tuned, leading to more reliable and precise 3D reconstructions from sparse input views.

3.3.2 Poses Global Alignment

Similar to DUSt3R, when using more than two input images, global pose alignment is needed to maintain consistency across views. This is because each image's predicted point map is generated at its own normalized scale, leading to misalignment in multi-view settings. To resolve this, we introduce a global optimization strategy that refines alignment across all images. Unlike pairwise alignment, our approach jointly optimizes transformation parameters, enforcing global consistency.

For each image pair $e = (n, m) \in \xi$, we consider the corresponding point maps $X^{n,n}$ and $X^{m,n}$, along with their associated confidence weights $C^{n,n}$ and $C^{m,n}$. Our objective is to transform all pairwise predictions into a unified coordinate system. To achieve this, we introduce a pairwise pose transformation T_e and a scaling factor σ_e. Thus, we redefine the global pose alignment as an optimization problem:

$$\hat{P}^* = \arg\min_{\hat{P},T,\sigma} \sum_{e\in\xi} \sum_{v\in e} \sum_{i=1}^{HW} C_i^{v,e} \left\| \hat{P}_i^v - \sigma_e T_e X_i^{v,e} \right\|. \tag{3.9}$$

Here, $v \in e$ implies that v takes values from the set $\{n, m\}$ when $e = (n, m)$. Given an image pair e, we optimize T_e to align both point maps $X^{v,e}$ with the world-coordinate point maps P^v. To prevent the trivial solution where all scaling factors collapse to zero (i.e., $\sigma_e = 0$ for all $e \in \xi$), we impose a constraint ensuring that their product remains unity: $\prod_{e\in\xi} \sigma_e = 1$. This constraint maintains a consistent scale across all transformations, enabling effective global pose alignment across multiple images.

3.3.3 Redundancy-Free Initialization

With globally aligned points, poses, and pixel-level confidence maps, we initially generate a dense point cloud. However, this introduces significant redundancy, where multiple points occupy the same spatial location, leading to inefficiencies in storage and computation. To address this, we propose a Redundancy-Free (RF) Strategy to downsample the point cloud, reducing redundancy while preserving essential scene details and improving representation efficiency.

Inspired by SLAM [26, 18], we randomly select a primary viewpoint and initialize new Gaussians using all its pixels. Since some regions are already well-represented, indiscriminately adding Gaussians would cause unnecessary duplication. To prevent this, we introduce a masking mechanism to selectively determine where new Gaussians should be introduced. The mask formulation is as follows:

$$M_p = (S_p < 0.5) + \left(D_p^{GT} < D_p\right) \left(L_1(D_p) > 50\,\text{MDE}\right). \tag{3.10}$$

Here, p represents each pixel, and S_p denotes the density of the Gaussian at that point, computed similarly to color and depth. The density function is defined as:

$$S_p = \sum_{i \in N} s_i \alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_j'), \tag{3.11}$$

where s_i represents the Gaussian weight, and α_i' is the opacity of the i-th Gaussian in the accumulation process. The mask M_p ensures that new Gaussians are only added in regions where the density is insufficient or where the estimated depth D_p is positioned in front of the ground truth depth D_p^{GT}, and the depth error exceeds 50 times the median depth error (MDE). The addition of new Gaussians follows the same initialization procedure as that used for the primary viewpoint.

3.4 Depth Regularization

In sparse-view settings, optimizing Gaussians solely with multi-view photometric loss is not that adequate, as it constrains appearance without fostering a coherent geometric structure. This limitation increases the risk of overfitting to training views and reduces generalization to test views, often causing floaters and ghosting. To address this issue, we introduce additional priors and regularization terms. Depth priors derived from pre-trained multi-view stereo will intuitively guide the model toward the correct geometric structure.

3.4.1 Training Depth Regularization

Our multi-view stereo depth prior originates from the pre-trained model DUSt3R [47], which provides relative depth. To address the scale ambiguity between real-world scenes and estimated depths, we employ a relaxed relative loss method, $Pearson$ correlation coefficient, formulated as follows:

$$\text{Corr}(D_{\text{ras}}, D_{\text{est}}) = \frac{\text{Cov}(D_{\text{ras}}, D_{\text{est}})}{\sqrt{\text{Var}(D_{\text{ras}})\text{Var}(D_{\text{est}})}}. \tag{3.12}$$

Here, D_{est} is the depth estimated by the multi-view stereo, and D_{ras} is the depth rendered by our system. This constraint benefits from being unaffected by scale inconsistencies, optimizing the correlation between the two depths.

3.4.2 Pseudo Depth Regularization

To improve the generalization of 3D Gaussian across unseen views and reduce overfitting, we generate pseudo views using Principal Component Analysis (PCA) [2] to refine camera poses and motion paths. We calculate focal points and derive elliptical path dimensions and altitude changes using statistical methods. Camera positions are interpolated along these paths using trigonometric functions. For each position, we compute the view matrix, orienting the camera towards the focal point and aligning the "up" vector with the average direction from the input views.

$$P_{\text{new}} = T^{-1} \cdot V_{(p,u,c)} \cdot T. \tag{3.13}$$

Here, P_{new} represents the newly generated camera pose matrix, T is the transformation matrix obtained through PCA that adjusts the original view data to an optimized coordinate system, $V_(p, u, c)$ is the view matrix function that computes the transformation from the world coordinate system to the camera coordinate system, and T^{-1} is the inverse of T, converting the adjusted pose matrix back to the original view coordinate system. The pseudo views generated are also subject to depth constraints, as follows:

$$\text{Corr}(D_{\text{ras}}^{pse}, D_{\text{est}}^{pse}) = \frac{\text{Cov}(D_{\text{ras}}^{pse}, D_{\text{est}}^{pse})}{\sqrt{\text{Var}(D_{\text{ras}}^{pse})\text{Var}(D_{\text{est}}^{pse})}}. \tag{3.14}$$

here, D_{ras}^{pse}represents the depth rendered from the pseudo viewpoint, while D_{est}^{pse} is the depth predicted by the depth pre-trained model based on the RGB image rendered from the same viewpoint.

3.5 Multi-view Appearance Refinement

Above, we improve 3D Gaussian representation by ensuring geometry consistency with added depth regularization. However, color inconsistencies persist in images generated from unseen views, leading to a noisy representation. To address this, we've developed a Multi-view Appearance Refinement (MAR) algorithm. It begins by rendering N images from the Gaussian field at predefined viewpoints, denoted as I_n for $n \in (1, N)$. These renderings are then refined using diffusion models to produce new images $\hat{I}_n$, which leverage photometric priors to optimize for correct color consistency.

3.5.1 Diffusion Process

Starting with the renderings I_n, we initially employ the forward process of the diffusion model to introduce noise, resulting in a set of noisy renderings, $\hat{I}_n^t$, where t represents a specific time step. Subsequently, we initiate the regeneration of these images by sampling from the Markov Chain in the reverse process, $\prod_{n=1}^{N}\prod_{t=1}^{T} p_\theta(\hat{I}_{t-1}^n|\hat{I}_t^n)$ while ensuring multi-view consistency among all images. To maintain consistency at every timestep, we train a 3D Gaussian field specifically to guide the denoising trajectory towards preserving this multi-view consistency. The predicting noise progress is :

$$\hat{\mu}(\hat{I}_t^n, t) = \frac{1}{\alpha_t}\left(\hat{I}_t^n - \beta_t \epsilon_0(\hat{I}_t^n, t)\right), \tag{3.15}$$

here, $\epsilon_0(\hat{I}_t^n, t)$ represents the predicted noise on the n-th rendered view at t timestep. The term $\hat{\mu}(\hat{I}_t^n, t)$ denotes the estimated $\hat{I}_t^n$, while α_t and β_t are predefined constants. The denoising step is as follows:

$$\hat{I}_{t-1}^n = s_t \hat{I}_t^n + d_t \hat{\mu}_t^n + \omega_t \gamma, \gamma \sim \mathcal{N}(0, I), \tag{3.16}$$

here, s_t, d_t and ω_t are all predefined constants, and γ represents noise sampled from a standard Gaussian distribution.

3.5.2 Pseudo Appearance Regularization

To introduce multi-view consistency, we train a 3DGS model without noise and then render images on it as $\mu^{'n}$. Then, the new one with consistency constrain is calculated as:

$$\tilde{\mu}_t^n = w_t \delta_t^n \mu_t^{'n} + (1 - w_t)\hat{\mu}_t^n, \tag{3.17}$$

here, δ_t represents the ratio of the standard deviation of $\hat{\mu_t}$ to that of $mu^{'n}$ to prevent overexposure. Meanwhile, w_t is a predefined weight used to balance the denoising outcomes.

After obtaining images from the pseudo viewpoints generated by the diffusion model, we use the loss of color the same as 3DGS to calculate the loss between the rendered images and the diffusion-generated images,

$$L_{\text{cp}} = L_1\left(\hat{I}, I^{'}\right) + \lambda^{'} L_{\text{D-SSIM}}\left(\hat{I}, I^{'}\right), \tag{3.18}$$

where $\hat{I}$ represents the rendering image, $I^{'}$ represents the diffusion-generated image.

3.6 Loss Function

As outlined above, the loss function consists of three main components: the color regularization loss L_c, the depth regularization loss for training views L_d, and the depth regularization loss for pseudo views L_{dp}. The overall loss function is formulated as follows:

$$L = \lambda_1 L_{\text{color}} + \lambda_2 L_{\text{d}} + \lambda_3 L_{\text{dp}} + \lambda_4 L_{\text{cp}}. \tag{3.19}$$

The color regularization loss L_{color}, computed based on the original 3D Gaussian representation, is defined as:

$$L_{\text{color}} = L_1(\hat{I}, I) + \lambda' L_{\text{D-SSIM}}(\hat{I}, I), \tag{3.20}$$

where $\hat{I}$ denotes the rendered image, and I represents the ground-truth image. Based on a grid search, we set the loss weighting parameters as follows: $\lambda_1 = 0.5, \quad \lambda_2 = 1, \quad \lambda_3 = 0.05, \quad \lambda_4 = 0.001$.. Additionally, following 3DGS [19], we set $\lambda' = 0.2$.

CHAPTER 4

Experiments

4.1 Experimental Setup

In this section, I will elaborate on the experimental setup, including the datasets used and the comparison methods. Additionally, I will present the experimental evaluation results and conduct comprehensive ablation studies to validate the effectiveness of each component.

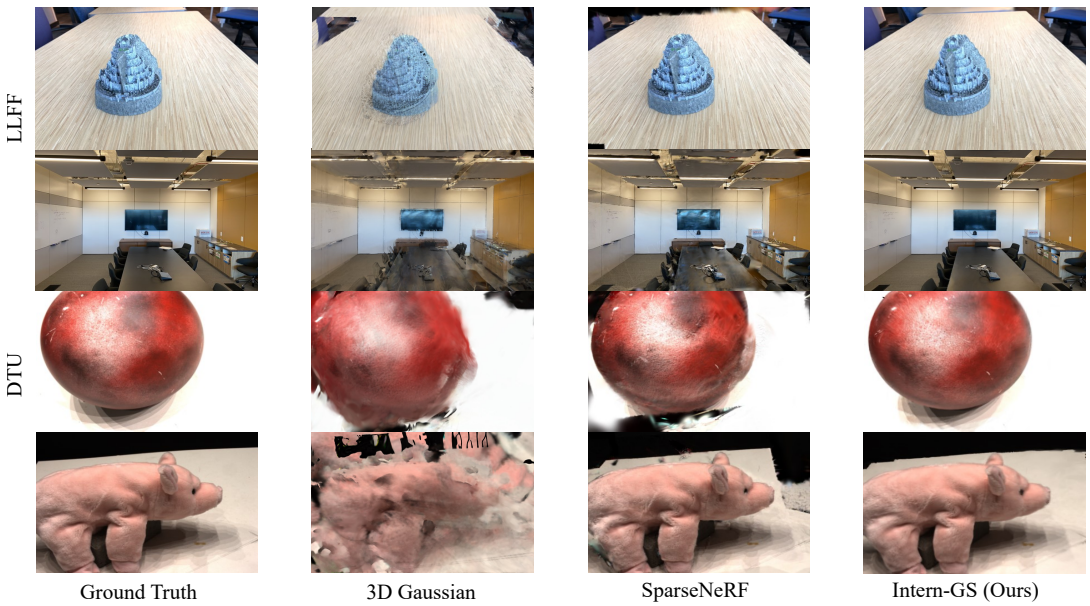

FIGURE 4.1. Results on LLFF dataset [28] and DTU dataset [1] in 3 training views. Our method captures more scene details, particularly in areas with sparse texture information. The 3D Gaussian [19] approach struggles to synthesize accurate new views under sparse viewpoints, while SparseNeRF [45] produces overly smooth views, losing many details.

FIGURE 4.2. Results on Tanks dataset [21] in 3 training views. In comparison, 3DGS [19] struggles to accurately represent structures. While SparseNeRF [45] performs well overall, it tends to lose some texture information in areas with flat depth. In contrast, Intern-GS effectively captures these texture details.

4.1.1 Datasets

We conduct our experiments on three widely used datasets, LLFF [28], DTU [1], and Tanks and Temples [21], to comprehensively evaluate our method across diverse scene types. For the LLFF dataset, we follow prior works [45, 54] by splitting the images into three designated training views and multiple test views. This setup ensures a fair comparison with existing methods while assessing our model's ability to generalize from sparse observations. For the DTU dataset, we adopt the experimental protocol used in SPARF [43] and RegNeRF [30], training our model on the same three training views and evaluating it on the corresponding test views. To eliminate background noise and focus solely on the target objects, we utilize the same object masks during evaluation, ensuring a consistent and controlled experimental setup. To further assess the model's applicability to non-forward-facing scenes, we conduct additional experiments on the Tanks and Temples dataset, following the approach outlined in InstantSplat [11]. This dataset presents more complex and diverse 3D environments, allowing us to evaluate our model's performance in challenging real-world scenarios. Regarding downsampling strategies, we apply downsampling rates of 8 and 4 for the LLFF and DTU datasets, respectively, to reduce computational overhead while maintaining sufficient detail. However, for the Tanks and Temples dataset, we use full-resolution images without any downsampling to preserve fine structural details and scene complexity.

TABLE 4.1. Comparison of PSNR, LPIPS [48], and SSIM [59] with current state-of-the-art methods for the novel view synthesis task on the LLFF [28] and DTU [1] datasets. All baseline results in the table are sourced from [45], and the state-of-the-art results are highlighted in bold black.

Methods	LLFF			DTU		
	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑
SRF [9]	12.34	0.591	0.250	15.32	0.304	0.671
PixelNeRF [56]	7.93	0.682	0.272	16.82	0.270	0.695
MVSNeRF [8]	17.25	0.356	0.557	18.63	0.197	0.769
Mip-NeRF [4]	14.62	0.495	0.351	8.68	0.353	0.571
DietNeRF [16]	14.94	0.496	0.370	11.85	0.314	0.633
RegNeRF [30]	19.08	0.336	0.587	18.89	0.190	0.695
FreeNeRF [54]	19.63	0.308	0.612	19.92	0.182	0.787
SparseNeRF [45]	19.86	0.328	0.624	19.55	0.201	0.769
3DGS [19]	15.52	0.405	0.408	10.99	0.313	0.585
InstantSplat [11]	17.67	0.379	0.603	17.55	0.212	0.634
Intern − GS(Ours)	**20.49**	**0.304**	**0.656**	**20.34**	**0.163**	**0.789**

4.1.2 Comparison Methods

Following prior work on neural radiance fields in the few-shot setting, we compare our Intern-GS with several state-of-the-art methods, including SRF [9], PixelNeRF [56], MVSNeRF [8], Mip-NeRF [4], DietNeRF [16], RegNeRF [30], FreeNeRF [54], SparseNeRF [45], and DS-NeRF [10].

Additionally, we report the results of raw 3D Gaussian Splatting [19] and the pose-free InstantSplat [11] for direct comparison. For fairness, the results of some previous works are taken directly from their respective published papers.

To quantitatively evaluate the reconstruction performance, we employ Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [59], and Learned Perceptual Image Patch Similarity (LPIPS) [48] as evaluation metrics across all methods, providing a comprehensive assessment of both accuracy and perceptual quality.

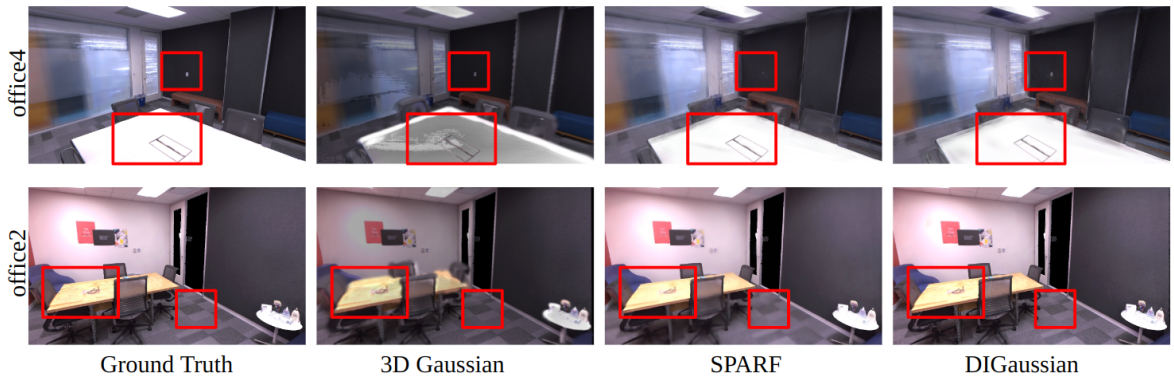

FIGURE 4.3. Results on Replica dataset [40] in 3 training views. In comparison, 3DGS [19] struggles to accurately represent structures. While SPARF [43] performs well overall, it tends to lose some texture information in areas with flat depth. In contrast, Intern-GS effectively captures these texture details.

4.1.3 Training Details

For the LLFF dataset [28], the original image size is 4032*3024. We use images that have been downsampled by a factor of eight to 504*378 as the input of DUSt3R [47]. Regarding the DTU dataset [1], the original image size is 1600*1200. We resize the image to 400*300 as the input of DUST3R. We did not do downsampling to Replica with size 1200*680. Subsequently, we will obtain a 3D pointmap of the same size. Next, we use global alignment to align these point maps to the same coordinate system. From this, we obtain consistent scale depth information and camera poses. Inspired by SLAM [26, 18], we use monocular depth information to initialize non-redundant 3D Gaussians according to a unique image sequence. Specifically, for the first image, we initialize all pixel points. For subsequent images, we only initialize areas with an uncertainty of less than 0.5. During all the 10000 training epochs, the learning rates for position, SH coefficients, opacity, scaling, and rotation are set to 0.00016, 0.0025, 0.05, 0.005, and 0.001, respectively. We begin with a Spherical Harmonics (SH) degree of 0 for basic lighting representation and increment it by one every 500 iterations until reaching a degree of 4, gradually increasing the complexity, and we reset the opacity for all Gaussians to 0.05 at iterations 2000, 5000, and 7000 aligned with the original 3D Gaussian splatting.

TABLE 4.2. Comparison of PSNR, LPIPS [48], and SSIM [59] with state-of-the-art (SOTA) methods on the Tanks and Temples dataset [21]. The results for RegNeRF and SparseNeRF are sourced from [45], while the results for 3DGS and InstantSplat are obtained from our own reproductions. The state-of-the-art results are highlighted in bold black.

Method	PSNR↑	LPIPS↓	SSIM↑
RegNeRF [30]	19.64	0.243	0.718
SparseNeRF [45]	21.98	0.219	0.730
3DGS [19]	15.36	0.379	0.572
InstantSplat [11]	22.20	0.199	**0.743**
Intern-GS (Ours)	**22.67**	**0.191**	0.736

4.2 Comparison to Baseline

In this section, we provide a comprehensive performance analysis of Intern-GS, our proposed algorithm, against several state-of-the-art methods across three datasets. The extensive evaluation aims to validate the efficacy of Intern-GS and assess its robustness under various noise settings.

4.2.1 Results on LLFF Dataset

Intern-GS has been comprehensively evaluated on the LLFF dataset, demonstrating robust performance across both qualitative and quantitative assessments. As shown in Figure 4.1 and Table 4.1, our method consistently outperforms all baseline approaches across all evaluation metrics. Statistically, Intern-GS achieves higher accuracy and fidelity, highlighting its effectiveness in reconstructing detailed and structurally consistent 3D representations.

From a visual perspective, our model excels in capturing fine-grained geometric details, particularly in challenging regions with sparse texture information. In contrast, 3DGS [19] encounters difficulties under sparse viewpoints, leading to incomplete or less reliable reconstructions. Although SparseGS [45] incorporates monocular depth as an additional constraint,

its reliance on sparse point cloud initialization results in lower geometric completeness compared to our method, which benefits from a dense initialization strategy. By leveraging a more robust depth-aware approach, Intern-GS effectively reconstructs complex structures with higher precision, making it particularly well-suited for real-world applications requiring high-quality 3D scene understanding.

4.2.2 Results on DTU Dataset

We present the results of our method on the DTU dataset in Figure 4.1 and Table 4.1. Unlike the LLFF dataset, the DTU dataset is characterized by a prominent central object against a black background. This distinction necessitates a careful evaluation process when computing metrics for novel view synthesis. Specifically, following the baseline approach, we apply a corresponding mask to each viewpoint and calculate evaluation metrics only within the masked image regions. This ensures that the black background does not interfere with the overall Gaussian optimization process, leading to a more accurate assessment of reconstruction quality.

Our results demonstrate that Intern-GS consistently achieves superior performance in terms of PSNR, LPIPS, and SSIM compared to previous methods. SparseNeRF, in particular, struggles with the dataset's noisy and smooth backgrounds, which adversely affect monocular depth estimation and, consequently, the quality of the reconstructed scene. Visually, our approach excels in handling low-texture and smooth regions, where it provides richer depth and color information. This enables our model to capture more precise structural and photometric details, significantly enhancing both the geometric fidelity and the overall visual realism of the synthesized views. These improvements underscore the robustness of our method in handling diverse scene characteristics, making it highly effective for complex real-world 3D reconstruction tasks.

TABLE 4.3. Ablation study on LLFF [28] with three training views, analyzing the individual contributions of the proposed three modules: Multi-View Stereo Guided Dense Initialization, Depth Regularization, and Multi-view Appearance Refinement. The results demonstrate that Multi-View Stereo Guided Dense Initialization has the most significant impact on the experimental outcomes.

Dense Init. (3.3)	Depth Regu. (4.3.2)	MAR (4.3.3)	PSNR ↑	LPIPS ↓	SSIM ↑
✗	✗	✗	15.52	0.405	0.408
✓	✗	✗	19.21	0.331	0.573
✓	✓	✗	19.64	0.329	0.640
✓	✓	✓	**20.49**	**0.304**	**0.656**

4.2.3 Results on Tanks and Temples Dataset

As shown in Table 4.2 and Figure 4.2, our Intern-GS achieved the best LPIPS and PSNR scores and second best in SSIM scores. This result showcases our model's capability with large scenes from non-frontal perspectives. For the unseen parts, our model is able to predict results with more consistent colors than SparseNeRF, which is attributable to our sophisticated diffusion prior.

4.3 Ablation Studies

We conduct ablation studies on three key aspects of our method: 4.3.1 a comparison between dense initialization and the conventional Structure-from-Motion-based initialization used in 3D Gaussian Splatting, 4.3.2 the impact of our proposed depth regularization on both training views and 4.3.3 novel synthesized views, and the effect of diffusion-prior-based appearance refinement in novel synthesized views. Table 4.3 presents the results of our ablation study on the LLFF dataset [28] using three training views as a case study.

4.3.1 Dense Initialization

We conducted a comparative analysis between our model with Dense and Non-redundancy Initialization and a baseline model without it. As shown in the first and second rows of Table 4.3, our approach consistently outperforms the original 3D Gaussian Splatting (3DGS) across all three evaluation metrics. This demonstrates that the Gaussian initialization generated via multi-view stereo provides a more reliable geometric prior, particularly in regions with sparse texture information. Furthermore, our redundancy reduction strategy effectively mitigates the influence of low-confidence areas, such as the sky, thereby enhancing the robustness of the initialization process.

4.3.2 Depth Regularization

We conducted a comparative study to evaluate the impact of our Depth Regularization by assessing models with and without it. As shown in the third row of Table 4.3, incorporating depth priors as constraints effectively guides Gaussian optimization toward more accurate geometric representations, enabling the model to learn more coherent and structurally faithful surfaces. This enhancement leads to a 0.43 increase in PSNR, a 0.002 reduction in LPIPS, and a 0.067 improvement in SSIM, demonstrating the effectiveness of our depth-aware regularization in refining scene reconstruction quality.

4.3.3 Multi-view Appearance Refinement

We conducted a comparative analysis to assess the impact of our Multi-view Appearance Refinement algorithm by evaluating models with and without it. In novel viewpoint regions, the absence of sufficient initialization points makes it challenging for Gaussians to be effectively optimized, leading to suboptimal representation of these areas. To address this, we introduce diffusion-based refinement from pseudo-views, enabling the model to better optimize from previously unseen perspectives. As shown in Table 4.3, the consistent improvement across all three evaluation metrics validates the effectiveness of our approach in enhancing appearance fidelity and overall reconstruction quality.

TABLE 4.4. Comparison (PSNR, LPIPS [48], SSIM [59]) with SOTA methods on Replica dataset [21].The results for RegNeRF and SparseNeRF are sourced from [45], while the results for 3DGS and InstantSplat are obtained from our own reproductions. The state-of-the-art results are highlighted in bold black.

Method	PSNR↑	LPIPS↓	SSIM↑
RegNeRF [30]	19.64	0.243	0.718
SparseNeRF [45]	21.98	0.219	0.730
3DGS [19]	15.36	0.379	0.572
InstantSplat [11]	22.20	0.199	**0.743**
Intern-GS (Ours)	**23.19**	**0.191**	0.736

4.4 Additional Experiments

In this section, we provide more experimental results to further demonstrate the effectiveness of our methods. To demonstrate that our approach is also applicable to synthetic indoor scenes, we evaluate the Replica dataset in Table 4.4 and Figure 4.3. In quantitative analysis, our method achieved state-of-the-art performance in LPIPS scores and demonstrated comparable performance in PSNR and SSIM scores. The slightly lower PSNR and SSIM scores compared to SPARF [43] can be attributed to SPARF's integration of pose optimization during training, which reduces the impact of pose noise on the model, which did not be considered in our framework. In qualitative analysis, 3DGS [19] can not represent the few-texture areas well; SPARF can effectively reconstruct scenes at the global level by incorporating additional depth priors, while Intern-GS captures more details in texture-scarce areas, particularly excelling in the continuous plane prediction of depth.

4.5 More Visual Results

In this section, we provide more visual results in continuous rendered views to demonstrate the superiority of the proposed method. The Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7 Shows the visual results on LLFF, DTU, Tanks and Temples datasets respectively.

FIGURE 4.4. The qualitative results of Intern-GS on LLFF dataset under 3 training views.

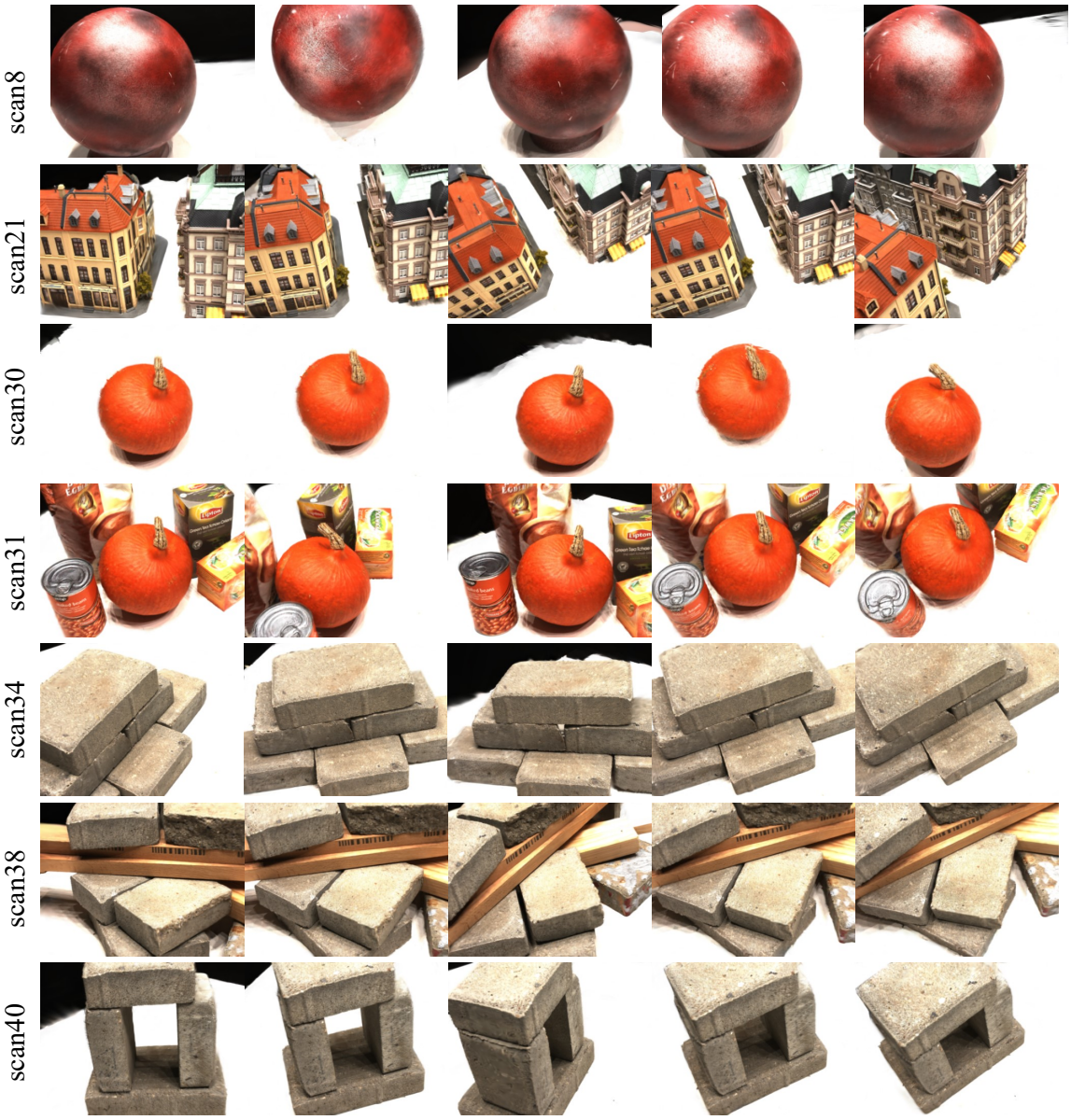

FIGURE 4.5. The qualitative results of Intern-GS on DTU dataset under 3 training views.

FIGURE 4.6. The qualitative results of Intern-GS on DTU dataset under 3 training views.

FIGURE 4.7. The qualitative results of Intern-GS on Tanks and Temples dataset under 3 training views.

CHAPTER 5

Discussion

5.1 Conclusion

In this work, we present Intern-GS, a novel view synthesis framework tailored for sparse-view scenarios, addressing the inherent limitations of traditional Structure-from-Motion (SfM)-based initialization. Our observations indicate that under sparse viewpoints, SfM tends to generate excessively sparse point clouds, which lack sufficient geometric and color priors, ultimately hindering the reconstruction process. This shortcoming is particularly problematic for 3D Gaussian Splatting, which heavily relies on a dense and well-informed initialization to achieve high-fidelity scene synthesis.

To tackle these challenges, Intern-GS leverages visual foundation models to predict both a richly informed point cloud initialization and more accurate camera parameters, providing a strong geometric prior from the outset. Additionally, our method integrates multi-view stereo priors to ensure effective and non-redundant initialization, particularly in texture-sparse regions where traditional methods struggle.

During Gaussian optimization, we further enhance reconstruction quality by introducing depth priors, which serve as constraints to guide the model toward more accurate and structurally coherent geometry. However, even with improved geometry, color inconsistencies and interpolation artifacts often arise due to gaps in initialization, especially in unseen regions. To address this, Intern-GS incorporates diffusion-based refinement from pseudo-views, effectively bridging missing color information and ensuring color-consistent optimization

under novel viewpoints. This approach enables the model to synthesize photorealistic and perceptually coherent renderings, even in regions that suffer from limited observations.

By seamlessly integrating geometry-aware initialization, depth regularization optimization, and diffusion-enhanced appearance refinement, Intern-GS provides a robust and comprehensive solution to the long-term challenge of rendering in texture-sparse environments. Our framework significantly improves view synthesis quality under sparse view constraints, paving the way for more accurate, consistent, and realistic 3D scene reconstructions.

5.2 Limitations and Future Directions

Although our method performs well on novel view synthesis tasks under sparse viewports, some limitations still exist. A key challenge is color inconsistency between different views, which is mainly caused by illumination changes. This problem is caused by the fact that the color representation of each 3D Gaussian in our model is view-independent and thus has difficulty adapting to illumination changes. In addition, under the constraint of dense initialization, the number of Gaussians contributing to the pixel color calculation is relatively limited, which further exacerbates the problem. To address this issue, we plan to explore view-dependent color modeling to provide more adaptive representations and enhance the realism of synthesized views.

In addition to improving color consistency, another interesting future research direction is to unify the generation and reconstruction paradigms in a single pipeline. Currently, reconstruction and generation are often regarded as two independent problems, but we believe that they can enhance each other. By reframing and interpreting these two tasks in a unified framework, we aim to develop a more cohesive and powerful approach to scene synthesis.

Bibliography

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola and Anders Bjorholm Dahl. 'Large-Scale Data for Multiple-View Stereopsis'. In: *International Journal of Computer Vision* (2016), pp. 1–16.

[2] Hervé Abdi and Lynne J Williams. 'Principal component analysis'. In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

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

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla and Pratul P Srinivasan. 'Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5855–5864.

[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*. 2023, pp. 4160–4169.

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis et al. 'Efficient geometry-aware 3d generative adversarial networks'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16123–16133.

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu and Hao Su. 'Tensorf: Tensorial radiance fields'. In: *European Conference on Computer Vision*. Springer. 2022, pp. 333–350.

[8] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu and Hao Su. 'Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14124–14133.

[9] Julian Chibane, Aayush Bansal, Verica Lazova and Gerard Pons-Moll. 'Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7911–7920.

[10] Kangle Deng, Andrew Liu, Jun-Yan Zhu and Deva Ramanan. 'Depth-supervised nerf: Fewer views and faster training for free'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12882–12891.

[11] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos et al. 'Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds'. In: *arXiv preprint arXiv:2403.20309* 2.3 (2024), p. 4.

[12] Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang and Ying He. '3d gaussian splatting as new era: A survey'. In: *IEEE Transactions on Visualization and Computer Graphics* (2024).

[13] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros and Xiaolong Wang. 'Colmap-free 3d gaussian splatting'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 20796–20805.

[14] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu and Jonathan Li. 'Nerf: Neural radiance field in 3d vision, a comprehensive review'. In: *arXiv preprint arXiv:2210.00379* (2022).

[15] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton and Julien Valentin. 'Fastnerf: High-fidelity neural rendering at 200fps'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14346–14355.

[16] 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*. 2021, pp. 5885–5894.

[17] 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*. 2021, pp. 5846–5854.

[18] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan and Jonathon Luiten. 'Splatam: Splat, track & map 3d gaussians for dense rgb-d slam'. In: *arXiv preprint arXiv:2312.02126* (2023).

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler and George Drettakis. '3d gaussian splatting for real-time radiance field rendering'. In: *ACM Transactions on Graphics* 42.4 (2023), pp. 1–14.

[20] Mijeong Kim, Seonguk Seo and Bohyung Han. 'Infonerf: Ray entropy minimization for few-shot neural volume rendering'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12912–12921.

[21] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou and Vladlen Koltun. 'Tanks and temples: Benchmarking large-scale scene reconstruction'. In: *ACM Transactions on Graphics (ToG)* 36.4 (2017), pp. 1–13.

[22] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba and Simon Lucey. 'Barf: Bundle-adjusting neural radiance fields'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 5741–5751.

[23] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua and Christian Theobalt. 'Neural sparse voxel fields'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15651–15663.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin and Baining Guo. 'Swin transformer: Hierarchical vision transformer using shifted windows'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[25] Zhiliang Ma and Shilong Liu. 'A review of 3D reconstruction techniques in civil engineering and their applications'. In: *Advanced Engineering Informatics* 37 (2018), pp. 163–174.

[26] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly and Andrew J Davison. 'Gaussian splatting slam'. In: *arXiv preprint arXiv:2312.06741* (2023).

[27] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He and Jingyi Yu. 'Gnerf: Gan-based neural radiance field without posed camera'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6351–6361.

[28] 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'. In: *ACM Transactions on Graphics (TOG)* 38.4 (2019), pp. 1–14.

[29] 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'. In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[30] 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'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5480–5490.

[31] Michael Niemeyer and Andreas Geiger. 'Giraffe: Representing scenes as compositional generative neural feature fields'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 11453–11464.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al. 'Learning transferable visual models from natural language supervision'. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[33] Christian Reiser, Songyou Peng, Yiyi Liao and Andreas Geiger. 'Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14335–14345.

[34] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers and Neil Houlsby. 'Scaling vision with sparse mixture of experts'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8583–8595.

[35] Antoni Rosinol, John J Leonard and Luca Carlone. 'Nerf-slam: Real-time dense monocular slam with neural radiance fields'. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 3437–3444.

[36] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. 'Learning representations by back-propagating errors'. In: *nature* 323.6088 (1986), pp. 533–536.

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

[38] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys and Jan-Michael Frahm. 'Pixelwise View Selection for Unstructured Multi-View Stereo'. In: *European Conference on Computer Vision (ECCV)*. 2016.

[39] Katja Schwarz, Yiyi Liao, Michael Niemeyer and Andreas Geiger. 'Graf: Generative radiance fields for 3d-aware image synthesis'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 20154–20166.

[40] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma et al. 'The Replica dataset: A digital replica of indoor spaces'. In: *arXiv preprint arXiv:1906.05797* (2019).

[41] Edgar Sucar, Shikun Liu, Joseph Ortiz and Andrew J Davison. 'imap: Implicit mapping and positioning in real-time'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6229–6238.

[42] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu and Gang Zeng. 'Dreamgaussian: Generative gaussian splatting for efficient 3d content creation'. In: *arXiv preprint arXiv:2309.16653* (2023).

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

[44] Shimon Ullman. 'The interpretation of structure from motion'. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 203.1153 (1979), pp. 405–426.

[45] Guangcong Wang, Zhaoxi Chen, Chen Change Loy and Ziwei Liu. 'Sparsenerf: Distilling depth ranking for few-shot novel view synthesis'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 9065–9076.

[46] Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong and Bisheng Yang. 'VistaDream: Sampling multiview consistent images for single-view scene reconstruction'. In: *arXiv preprint arXiv:2410.16892* (2024).

[47] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii and Jerome Revaud. 'DUSt3R: Geometric 3D Vision Made Easy'. In: *arXiv preprint arXiv:2312.14132* (2023).

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh and Eero P Simoncelli. 'Image quality assessment: from error visibility to structural similarity'. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[49] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen and Victor Adrian Prisacariu. 'NeRF–: Neural radiance fields without known camera parameters'. In: (2021).

[50] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan and Lin Gao. 'Recent advances in 3d gaussian splatting'. In: *Computational Visual Media* 10.4 (2024), pp. 613–642.

[51] Christopher Xie, Keunhong Park, Ricardo Martin-Brualla and Matthew Brown. 'Fig-nerf: Figure-ground neural radiance fields for 3d object category modelling'. In: *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, pp. 962–971.

[52] Haolin Xiong. *SparseGS: Real-time 360° sparse view synthesis using Gaussian splatting*. University of California, Los Angeles, 2024.

[53] Zhiwen Yan, Chen Li and Gim Hee Lee. 'Nerf-ds: Neural radiance fields for dynamic specular objects'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 8285–8295.

[54] 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*. 2023, pp. 8254–8263.

[55] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng and Angjoo Kanazawa. 'Plenoctrees for real-time rendering of neural radiance fields'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5752–5761.

[56] 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*. 2021, pp. 4578–4587.

[57] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan and Yonghong Tian. 'Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis'. In: *arXiv preprint arXiv:2409.02048* (2024).

[58] Kai Zhang, Gernot Riegler, Noah Snavely and Vladlen Koltun. 'Nerf++: Analyzing and improving neural radiance fields'. In: *arXiv preprint arXiv:2010.07492* (2020).

[59] 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*. 2018, pp. 586–595.

[60] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger and Andrew J Davison. 'In-place scene labelling and understanding with implicit scene representation'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 15838–15847.

[61] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik and Alexei A Efros. 'View synthesis by appearance flow'. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer. 2016, pp. 286–301.

[62] Fang Zhu, Shuai Guo, Li Song, Ke Xu, Jiayu Hu et al. 'Deep review and analysis of recent nerfs'. In: *APSIPA Transactions on Signal and Information Processing* 12.1 (2023).

[63] Zehao Zhu, Zhiwen Fan, Yifan Jiang and Zhangyang Wang. 'Fsgs: Real-time few-shot view synthesis using gaussian splatting'. In: *European conference on computer vision*. Springer. 2024, pp. 145–163.

[64] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald and Marc Pollefeys. 'Nice-slam: Neural implicit scalable encoding for slam'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12786–12796.