

Article

Large Scale Indoor Visual-Geometric Multimodal Dataset and Benchmark for Novel View Synthesis

Junming Cao ^{1,2†}, Xiting Zhao ^{3,†} and Sören Schwertfeger ^{3,*}

¹ Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China; caojm@sari.ac.cn

² University of Chinese Academy of Sciences, Beijing 100049, China;

³ The authors are with the Key Laboratory of Intelligent Perception and Human-Machine Collaboration – ShanghaiTech University, Ministry of Education, China; {zhaoxt, soerensch}@shanghaitech.edu.cn

* Correspondence: soerensch@shanghaitech.edu.cn

† These authors contributed equally to this work.

Abstract: Accurate reconstruction of indoor environments is crucial for applications in augmented reality, virtual reality, and robotics. However, existing indoor datasets are often limited in scale, lack ground-truth point clouds, and provide insufficient viewpoints, which impedes the development of robust novel view synthesis (NVS) techniques. To address these limitations, we introduce a new large-scale indoor dataset that features diverse and challenging scenes, including basements and long corridors. This dataset offers panoramic image sequences for comprehensive coverage, high-resolution point clouds, meshes, and textures as ground truth, and a novel benchmark specifically designed to evaluate NVS algorithms in complex indoor environments. Our dataset and benchmark aim to advance indoor scene reconstruction and facilitate the creation of more effective NVS solutions for real-world applications.

Keywords: Novel View Synthesis; 3D Reconstruction; Indoor Dataset; Benchmark

Citation: Cao, J.; Zhao, X.; Schwertfeger, S. Large Scale Indoor Visual-Geometric Multimodal Dataset and Benchmark for Novel View Synthesis. *Remote Sens.* **2024**, *1*, 0. <https://doi.org/>

Received:

Revised:

Accepted:

Published:

Copyright: © 2024 by the authors. Submitted to *Remote Sens.* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Background and Significance

Indoor environments, where people spend a significant portion of their day, are centers of economic activity and value creation. The task of indoor scene reconstruction plays a crucial role in computer vision by enabling the creation of digital replicas of physical spaces. These replicas are essential for numerous applications, including augmented reality (AR), virtual reality (VR), robotic simultaneous localization and mapping (SLAM), and human-environment interaction [1]. Accurate digital replication of physical environments is vital and holds immense significance in the digital era.

For example, in architecture and engineering, 3D modeling of indoor spaces can provide detailed insights into the internal structure and layout of buildings, leading to cost reductions and enhanced efficiency during construction phases. It can also offer critical information to emergency personnel for response and safety planning, optimize rescue plans, and reduce potential risks [2]. In the real estate sector, virtual tours enable remote viewing of properties, providing convenience to both buyers and sellers. In cultural heritage preservation, indoor 3D reconstruction allows for the digital conservation of historical buildings and offers immersive online experiences [3].

These diverse applications highlight the demand for accurate novel view synthesis, particularly in indoor settings. Achieving 3D reconstruction through the observation and computation of objects or scenes has been a long-standing task in computer vision. The evolution from traditional methodologies to neural networks has marked significant advancements in this field [4].

Traditional 3D reconstruction algorithms like COLMAP [5][6] have relied heavily on feature extraction and matching techniques, such as SIFT [7] or SURF [8] feature matching, followed by Structure from Motion (SfM) and Multi-View Stereo (MVS) for scene reconstruction. SfM involves extracting keypoints and descriptors from images, matching keypoints across images using descriptors, and estimating camera motion and feature point positions in space when sufficient matched feature points are available. The camera intrinsics and extrinsics, along with the sparse 3D point clouds, are recovered through triangulation and finely adjusted via optimization processes to minimize reprojection errors [5]. MVS, leveraging camera poses obtained through SfM, estimates depth maps for each perspective and fuses them into a consistent dense 3D point cloud [6]. Simultaneous Localization and Mapping (SLAM) is another technique used for reconstructing 3D environments and obtaining camera poses. SLAM is typically used for real-time applications where a robot or device must navigate and build a map simultaneously. In contrast, SfM or MVS processes are usually performed offline due to their computational intensity. Our dataset utilizes both Lidar-visual-inertial SLAM and SfM software for pose optimization, providing comprehensive ground truth data and ensuring high-quality reconstructions.

However, traditional methods only output a dense point cloud and camera poses, which pose challenges for synthesizing images from new perspectives. Consequently, researchers have explored neural network-based approaches for 3D reconstruction. In 2020, Neural Radiance Fields (NeRF) [9] emerged as a groundbreaking technique, enabling implicit reconstruction results after extensive computation. Despite NeRF's limitations, such as suboptimal performance, lengthy computation times, and slow rendering speeds, its introduction has significantly impacted the field of 3D reconstruction.

Subsequent works, including Instant-NGP [10] and TensoRF [11], have achieved geometric-level improvements in NeRF's training and rendering processes, enabling rapid training and high frame rates. Mip-NeRF [12] and its follow-up, Mip-NeRF 360 [13], have enhanced model expressiveness and rendering quality, particularly in unbounded scenes. Other variants, such as DS-NeRF [14], Point-NeRF [15], and Tetra-NeRF [16], have incorporated point clouds or depth supervision, improving performance in sparse perspectives and depth accuracy. Yet, none have achieved photorealistic precision in fine detail rendering for indoor settings.

In 2023, the introduction of 3D Gaussian Splatting [17] marked a significant milestone in the field of novel view synthesis. This model employs a semi-explicit representation, utilizing 3D Gaussian core spheres articulated through spherical harmonic functions to depict color and spatial information. Extensive engineering optimizations in sorting, rendering, and derivation allow for high-precision rendering at super-high frame rates. To further enhance the rendering effects of the 3D Gaussian model, Mip-Splatting [18] adds a filter to remove high-frequency artifacts, ensuring high image quality even after zooming in. SuGaR [19] proposes a method for extracting meshes from 3D Gaussian, which, by binding gaussians to the mesh surface, enhances rendering capabilities while achieving geometric extraction.

For the algorithms mentioned above, images and their corresponding camera poses are essential inputs for evaluating benchmark datasets. Traditionally, these camera poses are computed using SfM algorithms. Additionally, some NeRF-based algorithms and most 3D Gaussian-based algorithms require point clouds, which are typically generated by MVS algorithms. However, several challenges exist in this data collection pipeline for indoor scenes:

A primary challenge in indoor environments is the presence of repetitive textures and a scarcity of distinctive features. Traditional SfM algorithms heavily depend on identifying and matching features across images. In environments with repetitive patterns, such as wallpaper and floor tiles, feature matching algorithms can mistakenly identify similar textures as identical feature points across different locations. These ambiguities lead to inaccuracies or failures in camera pose estimation. Additionally, smooth surfaces like walls and furniture result in a lack of sufficient feature points, further complicating the extraction

of 3D information by SfM algorithms. Both the repetition of patterns and the scarcity of features can cause significant problems. This scarcity of features may also affect NeRF and other NVS algorithms, preventing them from converging properly and achieving accurate reconstructions.

Furthermore, the complex nature of indoor environments, combined with occlusions from furniture and other objects, necessitates capturing images from various angles to obtain a comprehensive view of the scene. This requirement for multi-view capture is particularly challenging to achieve with previous data collection pipelines that rely on handheld front-view mobile phones or DSLR cameras. Additionally, the sparse point clouds generated by SfM often struggle in indoor settings, resulting in low accuracy and high noise levels. These low-quality point clouds can significantly impact the performance of NVS algorithms, leading to inaccurate 3D reconstructions and degraded quality in synthesized novel views.

Recognizing the limitations of SfM alone for indoor environments, we have developed a specialized data collection platform to assemble a large-scale indoor dataset. This platform addresses these limitations and generates the high-quality input data needed for cutting-edge NVS algorithms. Our new dataset introduces the following key features:

Diverse and Challenging Indoor Environments: Our dataset surpasses typical room-scale scenes by including unique and complex indoor spaces. Basements, long corridors, and environments with intricate layouts and occlusions provide a rigorous testbed for NVS algorithms, pushing the boundaries of their capabilities.

High-Resolution, Multi-View Camera System: To capture the rich visual detail and geometric complexity of indoor spaces from as many angles as possible, we employ Insta360 consumer-grade and professional-grade panorama cameras. These cameras provide panoramic coverage, ensuring comprehensive visual information and facilitating robust reconstruction.

LiDAR Sensor: To obtain dense and accurate point clouds that overcome the limitations of SfM-derived point clouds, we integrate an Ouster OS0-128 LiDAR with 128 beams and a 90° vertical FOV into our platform. This setup allows us to generate dense point clouds with a density resolution of up to 5mm.

SLAM Integration: For precise camera pose estimation, especially in challenging, feature-sparse areas, we utilize a robust LiDAR-visual-inertial SLAM algorithm tailored specifically for indoor environments. This integration offers additional pose constraints, resulting in more accurate 3D reconstructions.

1.2. Related Work and Comparison

While datasets exist for both indoor scenes and large-scale environments, there is a crucial gap in datasets specifically designed for large-scale, indoor novel view synthesis. Shown in Table 1, indoor datasets are often room-scale with single camera views, whereas large-scale scene datasets are typically outdoor datasets. Let's examine the landscape:

1.2.1. Indoor Datasets

Scannet [20]: The ScanNet dataset collected 2.5 million perspectives containing over 1500 different indoor scenes, including apartments, offices, classrooms, bathrooms, and other indoor environments. The dataset provides RGB images, depth images, and corresponding intrinsic and extrinsic camera parameters. Additionally, ScanNet offers rich annotation information, including scene classification, individual labels, and 3D object bounding boxes. The room scale of the ScanNet dataset is small, and the image resolution is relatively low, at only 1296x968, with the depth image resolution at 640x480.

ScanNet++ [21]: ScanNet++ is a dataset similar to ScanNet, using DSLR and iPhone to capture high-resolution RGB images, and Faro Focus Premium to capture LiDAR data. Although the ScanNet++ dataset captures fewer scenes than ScanNet, it offers higher image resolution and more accurate mesh data. However, its scenes are also small-scale indoor rooms.

ARKitScenes [22]: ARKitScenes was captured by Apple using LiDAR and RGB cameras on an iPad, collecting more than 1900 scenes at 1920x1440 resolution. This is the largest indoor dataset and the accuracy is satisfactory. However, there is still the issue of small scale for individual scenes.

MuSHRoom [23]: The 'MuSHRoom' dataset addresses immersive real-time modeling using consumer-grade hardware for non-human perception and AR/VR. It offers room-scale 3D reconstruction and novel view synthesis data from multiple sensors (Azure Kinect, iPhone, and laser scanner). The dataset tackles challenges such as occlusion, motion blur, and illumination diversity, aiming to enhance 3D reconstruction and rendering."

1.2.2. Large-Scale Scene Datasets

Tank and Temples [24]: The Tank and Temples dataset is one of the earliest large-scale scene datasets, captured using high-resolution cameras. It includes camera poses reconstructed using SfM/MVS algorithms (COLMAP), with some scenes featuring true point clouds and their corresponding relationships. However, the dataset's indoor scenes are relatively small in scale, and the larger scenes with point clouds are all outdoors.

GigaMVS [25]: GigaMVS features scenes from the Summer Palace captured with an ultra-high-resolution camera, covering an area of over 40,000 square meters. However, this dataset is limited to outdoor scenes.

KITTI 360 [26]: KITTI-360 is a large-scale outdoor dataset designed for autonomous driving research. It includes omnidirectional RGB images and uses LiDAR and radar to capture surrounding point clouds. Additionally, it provides rich labels for various objects. Despite these features, the images in the KITTI 360 dataset are low in resolution, and the visible range is limited.

GauU-Scene [27]: The GauU-Scene V2 dataset combines highly accurate LiDAR point clouds with comprehensive RGB images. It surpasses existing datasets in both area and point count, covering over 6.5 square kilometers. The authors propose a novel method to align LiDAR and image data, addressing discrepancies in coordinate systems. They evaluated various reconstruction methods, highlighting limitations in current image-based metrics and emphasizing the importance of reliable ground truth for geometry reconstruction tasks.

1.2.3. Datasets for Novel View Synthesis

As mentioned earlier, many Novel View Synthesis (NVS) schemes are available, often tested in idealized settings, coming up with their own datasets. These include cameras arranged in hemispherical or cylindrical configurations, as seen with NeRF synthetic data or Mip-NeRF 360 data [13], or forward-facing views similar to LLFF [28]. Some autonomous driving research also uses datasets like KITTI-360 [26] for NVS, where images follow the direction of the road. Typically, these algorithms use small road segments for experiments. From a practical perspective, it is desirable for NVS to be applicable to large-scale indoor and outdoor scenes. Implicit representations are challenging to split and merge compared to explicit representations. However, block processing based on semi-explicit Gaussian representations is relatively simple. Despite the achievements in large-scale scene NVS, there is a lack of suitable public datasets for testing. Consequently, representative works such as Block-NeRF [29], Mega-NeRF [30] and VastGaussian [31] have collected their own datasets. Although their work is commendable, the lack of open-sourced datasets limits the development of other related work.

Currently, the field of SLAM (Simultaneous Localization and Mapping) has witnessed numerous mature Benchmark efforts [32] [33] [34]. Meanwhile, existing indoor datasets and benchmarks for Novel View Synthesis (NVS) predominantly suffer from limitations in scale and the absence of comprehensive ground truth data. Notably, they exhibit the following shortcomings:

Limited Scene Scale: Most available datasets feature small-scale scenes that do not adequately capture the complexity and diversity of real-world environments.

I/O	Dataset	Avg Size	Device	Resolution	Pose	GT
I	Scannet	22.6m ²	iPad Air2 Occipital Structure Sensor	1296x968	Hardware	Mesh
I	Scannet++	32.6m ²	Sony A7IV iPhone13 pro Faro laser scanner	1920x1440	COLMAP	Mesh Point
I	ARKitScenes	<200m ²	2020 iPad Pro Faro Focus S70	1920x1440	Hardware	Point
I	MuSHRoom	61.4m ²	Azure Kinect iPhone 12 Pro Max Faro Focus X130	1280x720 994x738	COLMAP	Mesh
O	Tanks & Temples	462m ²	DJI X5R Sony a7S II Faro Focus X330	3840x2160	COLMAP	Point
O	GIGAMVS	8667m ²	Nikon Camera Sony A7R4 Rigel VZ400	gigapixel	COLMAP	Point
O	KITTI360	73.7km	2x fisheye camera Velodyne HDL-64E SICK LMS 200	1440x1440	OXTS & RTK	Point
O	GauU-Scene	1.1km ²	DJI Zenmuse L1	5472x3648	COLMAP	Point
I	Ours	3145m ²	Insta360 Dual 1-Inch Insta360 Titan Ouster OS0-128	3072x3072 2972x5280	SLAM & SFM	Point Mesh

Table 1. Dataset Comparison. I/O denote Indoor or Outdoor dataset.

Scarcity of High-Quality Indoor Data: High-quality indoor datasets are rare, and those available do not offer high-resolution cameras or point clouds. Even when such data exists, it is typically derived from SLAM datasets, which are restricted to unidirectional paths and lack panoramic coverage.

Sparse Viewpoints: The provided viewpoints are often too few, which limits the ability to perform robust NVS. This is particularly problematic in indoor settings, where multiple angles are crucial for understanding the spatial layout.

Absence of Ground Truth Point Clouds: There is a significant lack of datasets providing ground truth point clouds, essential for accurate depth estimation and 3D reconstruction.

These limitations have significantly hindered progress in large-scale NVS and the application of indoor scene reconstruction. Many studies have been constrained to using small-sized datasets like ScanNet, which do not provide the necessary scope for advancing the field. In summary, existing datasets lack a comprehensive benchmark for large indoor scenes, which includes full-view clear images, accurate corresponding camera poses, and Lidar data. Therefore, we propose the IVGM dataset to address this gap.^D

This work introduces a comprehensive dataset and benchmark designed to address the aforementioned limitations, with the following contributions:

- We present the first large-scale indoor dataset featuring multiple diverse and challenging scenes, such as basements and long corridors, to better simulate real-world conditions. This makes it a robust tool for developing and testing novel view synthesis (NVS) algorithms, addressing the limitations of **limited scene scale** and **scarcity of high-quality indoor data**.
- Our dataset provides panoramic views captured by a multi-camera system, offering extensive coverage and facilitating complete scene understanding. This setup incorporates both consumer-grade and professional-grade panorama cameras, ensuring extensive coverage and high-quality visual data, overcoming the **sparse viewpoints** issue.

- We provide high-quality ground truth point clouds accompanied by detailed textured meshes, which enable precise depth perception and realistic NVS. The inclusion of both point clouds and meshes supports a wide range of NVS methods and applications, which compensates for the **absence of ground truth point clouds**.
220
221
222
223
224
225
226
227
- We introduce a thorough benchmark and conduct extensive experiments to evaluate the performance of NVS algorithms. This benchmark rigorously tests the efficacy of these algorithms in large-scale and complex indoor settings, providing valuable insights and driving advancements in the field.
228
229
230
231
232
233
234
235

2. IVGM Dataset

2.1. Collection Platform

To assemble our dataset, we employ a comprehensive set of sensors, detailed in Table 2 and Figure 1. These include the Ouster OS0-128 LiDAR, known for its wide field of view and high point density; the Insta360 ONE RS 1-Inch 360 Edition camera, which provides high-resolution 360-degree imagery; and the Insta360 Titan camera, capable of capturing professional-grade 11K panoramic videos. We have customized an electric Go-Kart to house these sensors, enabling efficient data collection in both indoor and outdoor environments.
236
237
238
239
240
241
242



Figure 1. Go-Kart data collection platform equipped with Ouster OS0-128 LiDAR and Insta360 cameras.

The Ouster OS0-128 LiDAR, combined with the Insta360 ONE RS 1-Inch 360 Edition camera, is adapted from the Stereye Polar 3D Scanner. This setup includes an Xsens Mti-630 IMU, which supports precise LiDAR-Inertial-Visual odometry, enabling real-time environmental scanning and previews on mobile devices. The Stereye Polar post-processing software streamlines the workflow by extracting essential data, including LiDAR and OneRS camera poses, colored point clouds, textured meshes, and raw RGB and LiDAR data. Synchronization between the Insta360 Titan camera and the OneRS camera is achieved
236
237
238
239
240
241
242

#	Sensor	Modalities	FOV	Resolution	FPS
1	Insta360 ONE RS 1-Inch	2x 1" Fisheye Camera	180°x180°	3072x3072	24
2	Ouster OS0-128	Spinning Lidar	90°x360°	128x1024	10
3	Insta360 Titan	8x M43 Fisheye Camera	180°x70°	2972x5280	5
4	Xsens Mt-630	IMU	-	-	400

Table 2. Sensors Specifications

through manual alignment using human clapping at the beginning and end of the video. The intrinsic and extrinsic parameters of the two cameras are calibrated using Kalibr [35]. We use the extrinsic parameters and the OneRS camera poses to get the Titan camera initial poses. To enhance the precision of the camera poses in large scale scenes even further, Agisoft Sfm software is utilized with the initial camera poses as a constraint. We may also use methods like CP+ [36] for camera pose refinement.

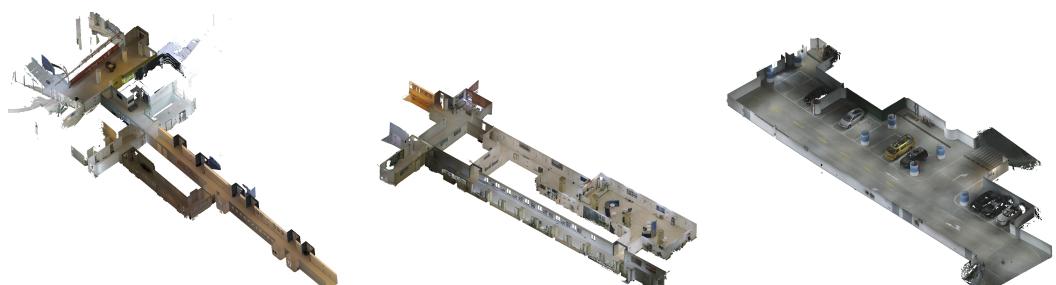
The Ouster OS0-128 REV6 is an ultra-wide field-of-view LiDAR sensor, featuring a 90° vertical FOV and a 35-meter range at 10% reflectivity. With 128 channels, it outputs up to 5.2 million points per second. Powered by the L2X digital LiDAR chip, this sensor can output up to 5.2 million points per second, capturing both the strongest and second-strongest light returns for each point [37].

The Insta360 ONE RS 1-Inch 360 Edition is a consumer-grade 6K 360 camera, featuring dual 1-inch Sony IMX283 rolling shutter sensors. These large sensors capture 3072x3072 fisheye images, enhancing light capture and detail for higher-quality images and videos with reduced noise [38].

The Insta360 Titan, on the other hand, is a professional-grade 360 camera that records 11K panoramic videos. It features eight Sony IMX269 Micro-4/3" sensors, known for their high image quality. These sensors, typically used in DSLR cameras, contribute to the Titan's exceptional image and video output [39].

2.2. Dataset Content

Sequence Name	Area Size(m ²)	Point Number	Insta Images	Titan Images
Office1	2,989.63	76,488,066	1,610	12,872
Office2	2,651.00	86,233,513	2,669	21,608
Underground	3,797.11	153,185,271	1,816	14,528

Table 3. Dataset Sequence Statistics

(a) Office Area 1

(b) Office Area 2

(c) Underground Garage

Figure 2. Three Dataset Sequence Groundtruth Mesh View

The IVGM dataset encompasses a diverse array of environments, meticulously captured by our custom-designed data acquisition vehicle across four distinct scenes. Shown in Table 3, this collection includes two segments from school office floors and one scene from underground garages.

Office Area 1: As depicted in Figure 2a, this area features long indoor corridors, a public zone with floor-to-ceiling windows overlooking an elevator, and a transparent glass platform offering a view of the atrium.

Office Area 2: Illustrated in Figure 2b, this area includes a series of indoor corridors, an outdoor corridor, a public rest area, administrative offices, and a designated tea break area for relaxation and informal interactions.

Underground Garage: The underground garage, shown in Figure 2c, is located under ShanghaiTech University. To achieve a higher density of visual perspectives, our data capture process involved multiple passes along the main thoroughfare. Additionally, to ensure optimal and uniformly distributed lighting, all lighting fixtures within the garage were activated during the capture process.

2.3. Challenges of IVGM Dataset

The IVGM dataset is captured in real-world conditions, presenting many complex situations not found in ideal datasets. These include significant exposure changes, reflections, transparent objects, motion blur, and a small number of dynamic objects.

Large Scene Area and High Perspective Density: Ideal test data for neural network visual reconstruction and novel view synthesis based on implicit representation typically involve small-scale, "object-centric" scenes with a central point of interest and uniformly distributed perspectives, using about 200 to 300 images. Each training image usually includes this central point, resulting in significant overlap between images. However, for large scenes, a central point does not exist, requiring the network to have robust expression and fitting capabilities. Moreover, large scenes contain a significant number of images, and storing point clouds requires substantial space. This demands advanced engineering capabilities to manage and process the extensive data efficiently.

Large Changes in Light: The IVGM dataset contains multiple real large-scale indoor scenes with complex lighting conditions. In Office Areas 1 and 2, sunlight shines through large glass windows. Although the data was collected during optimal times to minimize overexposure, significant differences between natural and artificial light are still evident in the images. In the Underground Parking Garage, all lights were turned on to ensure uniform brightness, though some areas remained dim. Learning-based NVS algorithms are highly sensitive to light variations, which are inevitable in large scenes. Thus, managing light changes is a significant challenge."

Motion Blur: The collection platform we designed can drive stably on smooth roads. In addition, we have limited the speed to 1m/s. However, due to changes in lighting, we can only set the shutter and white balance to automatic mode, which leads to a small amount of motion blur that is hard to avoid. This also poses a challenge to the learning ability of NVS algorithms.

Reflections and Transparent Objects: Indoor scenes often feature transparent and reflective surfaces, such as glass and metal. In Office Areas 1 and 2, windows, glass walls, and glass doors create transparency, while frames, elevators, and handrails add reflective properties. These elements pose challenges for both Lidar and camera on accurate data capture and novel view synthesis. Our previous work [40][41] addresses these challenges by developing methods for detecting reflections in 3D LiDAR scans. Incorporating these techniques into our current dataset may help mitigate the issues posed by reflective and transparent surfaces, leading to more accurate reconstructions and improved novel view synthesis.

3. Experiments

3.1. Evaluation of Algorithms

To evaluate the applicability, versatility, and performance of our dataset on novel view synthesis algorithms, we tested several popular methods developed in recent years. We categorized the algorithms into two groups: those requiring geometric information and those using only images as input.

Algorithms requiring only images as input: These algorithms are usually modifications based on NeRF.

NeRFStudio [42] offers researchers a robust algorithm framework, integrating many existing algorithms. Since its open-source release, many projects have been developed using its framework. The inputs for most algorithms within the NeRFStudio framework are standardized. Additionally, NeRFStudio introduces an extension algorithm: Nerfacto. Thus, we selected the NeRFStudio framework and its implemented algorithms for our tests.

Instant-NGP [10] introduces multi-resolution hash grid into NeRF and implements a lot of engineering optimization, significantly reducing the training time of NeRF to mere seconds.

Nerfacto [42]: Proposed by the author of NeRFStudio, Nerfacto integrates numerous previous works based on NeRF, altering the ray generation and sampling methods as well as the scene contraction method. Nerfacto offers three models of different scales: nerfacto, nerfacto-big and nerfacto-huge. We conducted experiments with nerfacto and nerfacto-huge models.

Algorithms requiring geometric input: These algorithms typically use depth images or point clouds as geometric priors. We selected two representative algorithms.

READ [43] which takes images and point clouds as inputs and proposes a rendering network. Using Monte Carlo sampling and multi-scale feature fusion, it performs novel view synthesis based on autonomous driving tasks.

3D Gaussian Splatting [17] uses images and point clouds as inputs, employing them as the basis for 3D Gaussian spheres. It utilizes spherical harmonics to represent anisotropic colors and applies backpropagation to optimize their position, color and shape.

3.2. Data Preprocessing

Common NVS algorithms and datasets typically employ consumer-grade smartphones or cameras to record from a single viewpoint. To simulate this scenario, we selected only the front fisheye camera for the consumer-grade Insta360 One R camera. Most current NVS algorithms primarily support the pinhole camera model. The two types of image data provided in our dataset are both fisheye and distorted. To make our dataset more relevant and easily applicable to the majority of existing NVS methodologies, we processed them as follows to create three types of data:

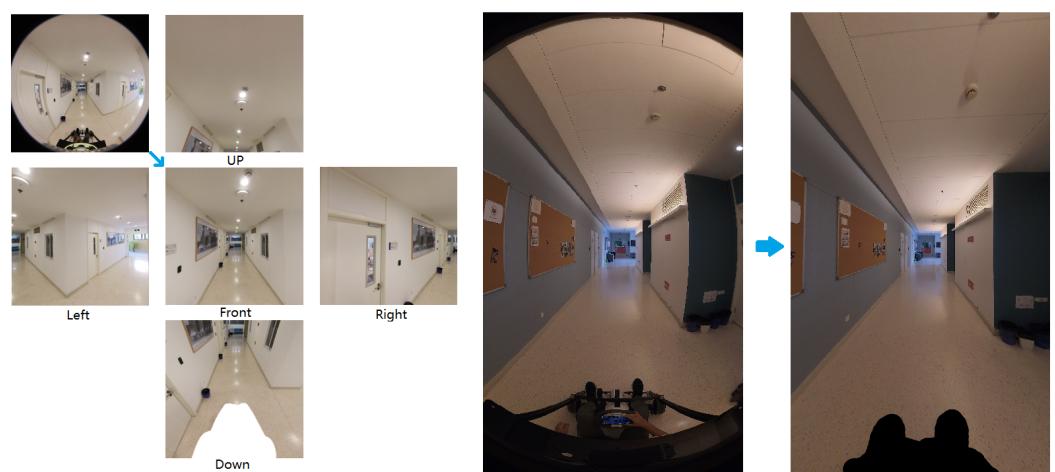


Figure 3. Comparison of original and processed images for fisheye undistortion using Insta360 One R(left) and Insta360 Titan(right) cameras.

Insta OneRS Camera Single View: We simply undistorted the fisheye images using the OpenCV fisheye model to pinhole image. The images were undistorted to a width and height of 1000 pixels each, with a field of view of 90 degrees. This approach crops out the surrounding perspective distortion while retaining more scene content.

Insta OneRS Camera Five View: The Insta OneRS camera has fisheye lens that capture a wider field of view than pinhole lens, providing more angular information. Directly undistorting the fisheye image to single view can lead to substantial loss of captured details. To address this, we have adopted a similar approach as outlined in the paper [44]. We projected the fisheye image onto a virtual hemisphere, then onto a virtual pinhole camera. We set up five virtual cameras facing up, down, left, right, and front, respectively. The width, height, and field of view of each camera match the previous fisheye undistort settings, so the image of the forward camera is identical to the single view undistorted result.

Insta Titan camera: There is only a small amount of wide-angle distortion in the images of the Insta Titan camera, so we chose to undistort the images using the fisheye model. The processed images have a width of 999 pixels, a height of 1776 pixels, and a horizontal field of view of 90 degrees.

In some images, the devices used to capture the data are visible, so we manually added masks to these images. Figure 3 shows the result of original and processed images. We also preprocessed the collected point clouds, downsampling them to a spacing of 3 cm to reduce GPU memory consumption.

3.3. Implementation Details

All training hyper-parameters followed the original paper's settings in our experiments. Each scene was trained on a single Nvidia 3090 GPU for approximately 3-22 hours, depending on the method's time complexity, with 4x Nvidia 3090 GPUs used in parallel. Due to the large number of images in the large scenes, we have 600 GB RAM per device.

For the Nerfstudio methods, we trained without modifying the parameters. For READ, based on the official website guidance, we changed the `-crop_size` to 448×448 according to our memory capacity. For 3D Gaussian Splatting, following the author's instructions for large dataset, we set `-position_lr_init` to 0.000016 and `-scaling_lr` to 0.001 , and increased the number of iterations to 20-30 times the number of images. All methods enabled the image mask.

3.4. Evaluation Metrics

To evaluate the performance of each method, we used three common metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [45], and Learned Perceptual Image Patch Similarity (LPIPS) [46] for novel view synthesis. The interpretation of these metrics is straightforward: higher PSNR and SSIM scores suggest superior image quality, whereas a lower LPIPS score denotes a more accurate perceptual match to the original scene.

PSNR: Peak Signal-to-Noise Ratio measures the quality of reconstructed images. It is calculated by averaging the squared pixel differences between the original image and the reconstructed image, followed by taking the logarithm.

SSIM: Structural Similarity assesses the structural similarity between two images. It considers luminance, contrast, and structural information. Higher SSIM values indicate better perceptual similarity.

LPIPS: Learned Perceptual Image Patch Similarity quantifies the perceptual difference between images. It leverages deep neural networks to capture perceptual features. Lower LPIPS scores indicate better perceptual similarity.

These metrics provide valuable insights into the quality and perceptual fidelity of synthesized views, aiding in the evaluation and comparison of different methods.

4. Benchmark

4.1. Benchmark for Different Algorithm on Different Sensors

Table 4 and Figure 4 show the results of various novel view synthesis (NVS) algorithms when applied to images captured by different sensors.

The Gaussian Splatting method consistently outperforms others across all scenes, as evidenced by the highest PSNR and SSIM values and the lowest LPIPS scores. This highlights the effectiveness of the semi-explicit representation used by Gaussian Splatting in rendering high-quality images in complex indoor environments. Instant NGP achieves commendable results, particularly with the Insta360 Titan camera, showing significant improvements in PSNR and SSIM when higher resolution images are used. For both Nerfacto and Nerfacto-huge, performance improves with higher resolution images from the Titan camera, although Gaussian Splatting still leads overall. Nerfacto-huge, despite its larger model size, performs slightly worse than the smaller Nerfacto model, potentially due to the difficulty in converging larger models in complex indoor scenarios. READ exhibits variable results, performing well in certain scenarios but not consistently across all camera settings.

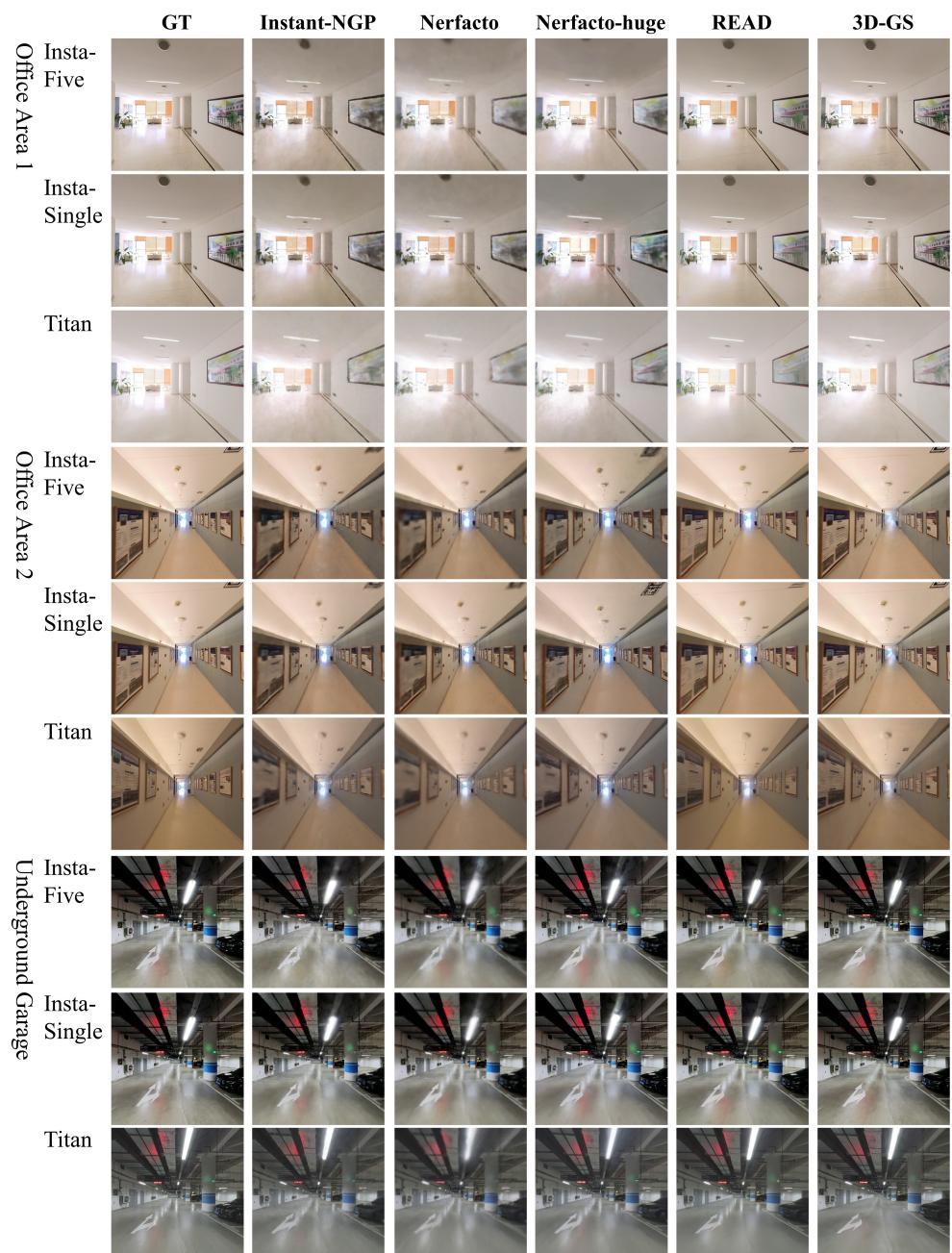


Figure 4. Qualitative Comparison of Different Methods on IVGM Dataset. The Titan result images are cropped to Insta images size.

Method	Office Area 1			Office Area 2			Underground Garage			
	Single	Five	Titan	Single	Five	Titan	Single	Five	Titan	
Instant NGP	PSNR↑	22.35	22.68	24.66	25.85	25.97	26.69	24.90	23.67	28.69
	SSIM↑	0.840	0.821	0.885	0.901	0.788	0.910	0.804	0.753	0.907
	LPIPS↓	0.334	0.352	0.299	0.308	0.416	0.318	0.378	0.429	0.384
Nerfacto	PSNR↑	19.48	20.22	24.07	23.67	23.01	25.76	24.35	23.66	28.10
	SSIM↑	0.782	0.784	0.874	0.873	0.850	0.901	0.779	0.747	0.900
	LPIPS↓	0.368	0.375	0.308	0.330	0.359	0.327	0.412	0.453	0.403
Nerfacto huge	PSNR↑	18.12	18.72	21.65	21.06	20.79	23.89	23.37	23.13	25.73
	SSIM↑	0.758	0.766	0.859	0.843	0.829	0.891	0.797	0.775	0.893
	LPIPS↓	0.379	0.377	0.313	0.345	0.369	0.336	0.340	0.389	0.362
READ	PSNR↑	27.12	19.07	22.38	27.75	19.99	20.48	27.63	22.68	22.89
	SSIM↑	0.873	0.783	0.866	0.884	0.789	0.835	0.845	0.732	0.861
	LPIPS↓	0.212	0.383	0.285	0.239	0.382	0.338	0.264	0.362	0.357
Gaussian Splatting	PSNR↑	28.42	28.09	28.84	31.32	32.27	31.27	28.04	29.87	30.00
	SSIM↑	0.899	0.870	0.909	0.912	0.938	0.930	0.862	0.892	0.921
	LPIPS↓	0.242	0.282	0.253	0.280	0.235	0.277	0.203	0.151	0.223

Table 4. Quantitative Comparison of Different Novel View Synthesis Methods in IVGM Dataset. The results for each scene are independently ranked and the top two values are highlighted. Red is the highest and Orange is second highest.

The results also indicate that the choice of camera system can significantly affect the performance of NVS algorithms. For instance higher resolution and multi view sensors, such as the Insta360 Titan camera, generally enhance the performance of NVS algorithms, highlighting the importance of image quality in these applications.

4.2. Cross Camera Evaluation

Insta One Camera Single View vs Five View

Method	Camera Path	Office Area 1		Office Area 2		Underground Garage	
		Single	Five	Single	Five	Single	Five
	Training Set	Single	Five	Single	Five	Single	Five
Instant NGP	PSNR↑	21.41	22.87	23.59	26.39	22.75	23.99
	SSIM↑	0.813	0.837	0.869	0.898	0.751	0.778
	LPIPS↓	0.357	0.346	0.346	0.317	0.425	0.402
Nerfacto	PSNR↑	19.26	20.19	22.44	23.43	22.77	23.96
	SSIM↑	0.770	0.789	0.845	0.866	0.739	0.765
	LPIPS↓	0.379	0.379	0.358	0.349	0.445	0.432
Nerfacto huge	PSNR↑	18.05	18.81	20.46	21.25	21.96	23.30
	SSIM↑	0.747	0.769	0.819	0.845	0.752	0.791
	LPIPS↓	0.390	0.375	0.371	0.355	0.399	0.353
READ	PSNR↑	24.00	18.99	23.93	20.34	24.19	23.41
	SSIM↑	0.826	0.806	0.844	0.806	0.780	0.758
	LPIPS↓	0.333	0.333	0.326	0.348	0.351	0.321
Gaussian Splatting	PSNR↑	23.34	27.50	26.17	31.28	23.03	27.19
	SSIM↑	0.826	0.869	0.874	0.930	0.753	0.859
	LPIPS↓	0.316	0.260	0.318	0.244	0.235	0.178

Table 5. Quantitative Results of Different Input Image Dataset on IVGM Dataset. Images of Insta-single results are rendered using Insta-five camera pose and vice versa.

Table 5 presents a cross-camera evaluation of NVS algorithms using a single image versus five images from the Insta360 ONE RS 1-Inch 360 Edition camera. Novel View Synthesis (NVS) algorithms strive to fit the training images as closely as possible. Consequently, when only the frontal view is provided, images from other directions (left, right, up, and down) lack a default perspective and are typically excluded from evaluation. However, the

five-view setup includes the frontal perspective, the comparison demonstrating the impact of multiple viewpoints on algorithm performance.

The evaluation shows that models trained using images from the frontal camera render noticeably lower quality results on the five-camera view compared to rendering on their respective training sets. In contrast, models trained with images from all five cameras better fit the images from all perspectives. The render results of the front view are not only similar to all views but also significantly better than those from single-camera training. This indicates that synthesizing views from unobserved angles, or free-view synthesis, can severely impact rendering quality. Our five-view preprocessing method is effective and improves overall performance.

Insta Camera vs Titan Camera

The Insta360 ONE RS (Insta) and Insta360 Titan (Titan) cameras are both oriented towards the front, each with a horizontal field of view (FOV) of 90 degrees. However, the Insta camera is positioned approximately 60 cm above the Titan camera. To compare the frontal view rendering effects of the two cameras, we rendered images from each camera in each other's pose using the Gaussian Splatting method. Figure 5 shows that the Titan camera, which captures a larger number of images and has a wider field of view, significantly reduces noise in the reconstructed results. Additionally, the varying color tones among the different camera groups provide a rich dataset for future research in multi-camera style transfer.



Figure 5. Qualitative Comparison of Gaussian Splatting Render on Insta and Titan Camera Pose.

4.3. Point Cloud Effect

Data	LiDAR-3cm			LiDAR-9cm			SfM		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Office Area 1	28.09	0.870	0.282	26.40	0.850	0.392	24.08	0.846	0.387
Office Area 2	32.27	0.938	0.235	23.12	0.842	0.407	18.03	0.791	0.372
Underground Garage	29.87	0.892	0.151	25.86	0.822	0.393	17.78	0.681	0.485

Table 6. Quantitative Results of Gaussian Splatting with Different Input Point Clouds.

We applied the Gaussian Splatting algorithm to sparse Structure-from-Motion (SfM) point cloud and LiDAR point cloud data which are downsampled to spacing of 3 cm and 9 cm respectively. Insta-five images are used for all settings and training iterations are set to 200,000. Figure 6 and Table 6 illustrate the impact of integrating LiDAR point cloud data with images using the Gaussian Splatting algorithm. The inclusion of LiDAR point cloud data significantly enhances the performance of the novel view synthesis (NVS)

algorithm, resulting in higher PSNR and SSIM values and lower LPIPS, compared to using only image data or sparse point clouds from SfM. In addition, denser pointclouds also boost performance. These results indicate that high-quality point cloud data, such as that from LiDAR, is crucial for improving the accuracy and quality of novel view synthesis, especially in complex indoor environments.

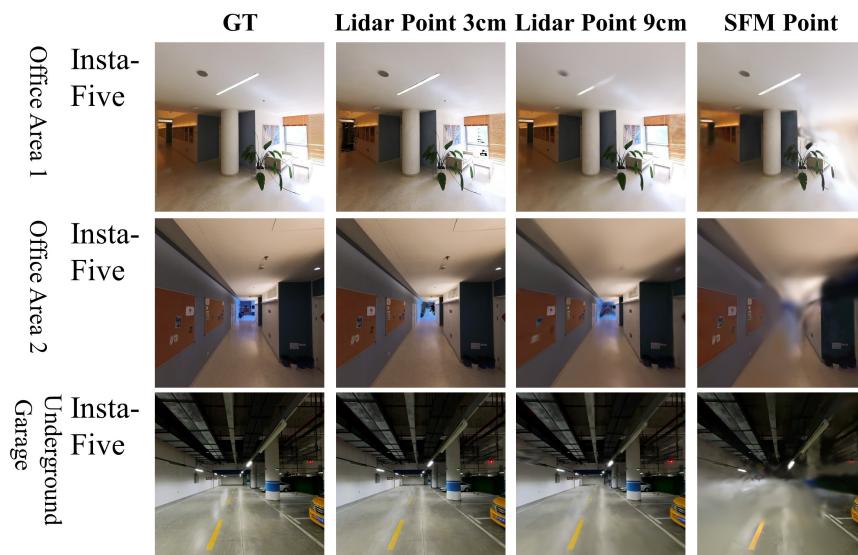


Figure 6. Qualitative Comparison of Gaussian Splatting with Different Input Point Clouds.

5. Discussion

The IVGM dataset represents a significant advancement in the field of novel view synthesis (NVS) and indoor scene reconstruction. By addressing the limitations of existing datasets, it provides a robust and comprehensive benchmark for large-scale indoor environments. A notable contribution of this dataset is its inclusion of diverse and challenging scenes, such as basements and long corridors, which offer a more accurate reflection of real-world conditions than the often idealized settings of previous datasets.

The use of a multi-camera system to capture panoramic views significantly enhances the dataset's utility by ensuring extensive coverage and facilitating a comprehensive understanding of the scene, which is vital for NVS algorithms. The dataset includes images from both consumer-grade and professional-grade cameras, supporting a wide array of research applications and ensuring versatility.

Reconstruction results obtained with the industrial-grade Titan camera outperform those achieved with the consumer-grade Insta One RS camera, mainly due to the Titan's broader field of view, lower noise levels, and higher image quality. Despite this, the Titan camera's higher tolerance for exposure bias makes it prone to overexposure—a drawback for image-based reconstruction. Future research may focus on harmonizing the exposure and color profiles of both cameras to enable simultaneous input of images from both, facilitating more comprehensive reconstructions. The experiment also demonstrates that undistorting large FOV fisheye images into five separate images retains more information and results in better reconstruction quality. Although this five-view undistortion may also lead to some information loss, researchers should explore direct training on fisheye models to achieve better results.

The high-quality ground truth point clouds, detailed meshes, and textures provided by the IVGM dataset are significant strengths, enabling precise depth perception and realistic NVS, crucial for applications in robotics, virtual reality, and augmented reality. However, methods like Gaussian Splatting have not fully utilized the high-quality LiDAR point

clouds, using them only for initialization. Better utilization of these high-quality point clouds could enhance rendering quality significantly.

6. Conclusions

We introduce a novel data acquisition device capable of simultaneously capturing high-precision LiDAR point clouds and two sets of high-resolution images. This device provides accurate camera parameters for the captured images. Using this equipment, we have collected a diverse array of large-scale indoor scenes within a campus environment, including office spaces and underground garages, resulting in the first large-scale indoor campus environment dataset.

The IVGM dataset sets a new standard for indoor visual-geometric datasets by addressing significant gaps in scale, quality, and comprehensiveness present in existing datasets. Its inclusion of large-scale, diverse, and challenging indoor scenes, coupled with high-resolution panoramic imagery and detailed ground truth data, makes it an invaluable resource for advancing the fields of NVS and indoor scene reconstruction.

We have also established a benchmark for indoor reconstruction and conducted experiments with several state-of-the-art novel view synthesis methods. Our findings indicate that existing methods can successfully perform reconstruction and new view synthesis on our dataset, with high-precision point clouds enhancing the quality of results for methods relying on point clouds as input. We anticipate that our dataset will serve as a valuable resource for future research in large-scale indoor reconstruction and generation.

In conclusion, the IVGM dataset addresses the critical need for a comprehensive indoor visual-geometric dataset and benchmark, paving the way for future research and development in novel view synthesis and beyond.

Author Contributions: Conceptualization, J.C. and X.Z.; methodology, X.Z.; software, J.C.; validation, J.C. and X.Z.; formal analysis, J.C.; investigation, X.Z.; resources, X.Z.; data curation, X.Z.; writing—original draft preparation, J.C. and X.Z.; writing—review and editing, J.C. and X.Z.; visualization, J.C.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Commission of Shanghai Municipality (STCSM), project 22JC1410700 "Evaluation of real-time localization and mapping algorithms for intelligent robots". This work has also been partially funded by the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence.

Data Availability Statement: The sensor datasets and ground truth data utilized in this work will be available online at <http://ivgm-dataset.github.io/>.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kang, Z.; Yang, J.; Yang, Z.; Cheng, S. A review of techniques for 3d reconstruction of indoor environments. *ISPRS International Journal of Geo-Information* **2020**, *9*, 330.
2. Díaz-Vilariño, L.; Khoshelham, K.; Martínez-Sánchez, J.; Arias, P. 3D modeling of building indoor spaces and closed doors from imagery and point clouds. *Sensors* **2015**, *15*, 3491–3512.
3. Spittle, B.; Frutos-Pascual, M.; Creed, C.; Williams, I. A review of interaction techniques for immersive environments. *IEEE Transactions on Visualization and Computer Graphics* **2022**.
4. Pintore, G.; Mura, C.; Ganovelli, F.; Fuentes-Perez, L.; Pajarola, R.; Gobbetti, E. State-of-the-art in automatic 3D reconstruction of structured indoor environments. In Proceedings of the Computer Graphics Forum. Wiley Online Library, 2020, Vol. 39, pp. 667–699.
5. Schönberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
6. Schönberger, J.L.; Zheng, E.; Pollefeys, M.; Frahm, J.M. Pixelwise View Selection for Unstructured Multi-View Stereo. In Proceedings of the European Conference on Computer Vision (ECCV), 2016.
7. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **2004**, *60*, 91–110.
8. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In Proceedings of the Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part I 9. Springer, 2006, pp. 404–417.

9. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **2021**, *65*, 99–106. 538
539
10. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)* **2022**, *41*, 1–15. 540
541
11. Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; Su, H. Tensorf: Tensorial radiance fields. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 333–350. 542
543
12. Barron, J.T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P.P. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5855–5864. 544
545
13. Barron, J.T.; Mildenhall, B.; Verbin, D.; Srinivasan, P.P.; Hedman, P. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5470–5479. 547
548
549
14. Deng, K.; Liu, A.; Zhu, J.Y.; Ramanan, D. Depth-supervised nerf: Fewer views and faster training for free. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12882–12891. 550
551
15. Xu, Q.; Xu, Z.; Philip, J.; Bi, S.; Shu, Z.; Sunkavalli, K.; Neumann, U. Point-nerf: Point-based neural radiance fields. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5438–5448. 552
553
16. Kulhanek, J.; Sattler, T. Tetra-nerf: Representing neural radiance fields using tetrahedra. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 18458–18469. 554
555
17. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* **2023**, *42*, 1–14. 556
557
18. Yu, Z.; Chen, A.; Huang, B.; Sattler, T.; Geiger, A. Mip-splatting: Alias-free 3d gaussian splatting. *arXiv preprint arXiv:2311.16493* **2023**. 558
559
19. Guédon, A.; Lepetit, V. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *arXiv preprint arXiv:2311.12775* **2023**. 560
561
20. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5828–5839. 562
563
564
21. Yeshwanth, C.; Liu, Y.C.; Nießner, M.; Dai, A. Scannet++: A high-fidelity dataset of 3d indoor scenes. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12–22. 565
566
22. Baruch, G.; Chen, Z.; Dehghan, A.; Dimry, T.; Feigin, Y.; Fu, P.; Gebauer, T.; Joffe, B.; Kurz, D.; Schwartz, A.; et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897* **2021**. 567
568
23. Ren, X.; Wang, W.; Cai, D.; Tuominen, T.; Kannala, J.; Rahtu, E. MuSHRoom: Multi-Sensor Hybrid Room Dataset for Joint 3D Reconstruction and Novel View Synthesis. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 4508–4517. 569
570
571
24. Knapitsch, A.; Park, J.; Zhou, Q.Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)* **2017**, *36*, 1–13. 572
573
25. Zhang, J.; Zhang, J.; Mao, S.; Ji, M.; Wang, G.; Chen, Z.; Zhang, T.; Yuan, X.; Dai, Q.; Fang, L. GigaMVS: a benchmark for ultra-large-scale gigapixel-level 3D reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *44*, 7534–7550. 574
575
26. Liao, Y.; Xie, J.; Geiger, A. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 3292–3310. 576
577
27. Xiong, B.; Zheng, N.; Li, Z. GauU-Scene V2: Expanse Lidar Image Dataset Shows Unreliable Geometric Reconstruction Using Gaussian Splatting and NeRF. *arXiv preprint arXiv:2404.04880* **2024**. 578
579
28. Mildenhall, B.; Srinivasan, P.P.; Ortiz-Cayon, R.; Kalantari, N.K.; Ramamoorthi, R.; Ng, R.; Kar, A. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)* **2019**, *38*, 1–14. 580
581
29. Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P.P.; Barron, J.T.; Kretzschmar, H. Block-nerf: Scalable large scene neural view synthesis. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8248–8258. 582
583
584
30. Turki, H.; Ramanan, D.; Satyanarayanan, M. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12922–12931. 585
586
31. Lin, J.; Li, Z.; Tang, X.; Liu, J.; Liu, S.; Liu, J.; Lu, Y.; Wu, X.; Xu, S.; Yan, Y.; et al. VastGaussian: Vast 3D Gaussians for Large Scene Reconstruction. *arXiv preprint arXiv:2402.17427* **2024**. 587
588
32. Nardi, L.; Bodin, B.; Zia, M.Z.; Mawer, J.; Nisbet, A.; Kelly, P.H.; Davison, A.J.; Luján, M.; O’Boyle, M.F.; Riley, G.; et al. Introducing SLAMBench, a performance and accuracy benchmarking methodology for SLAM. In Proceedings of the 2015 IEEE international conference on robotics and automation (ICRA). IEEE, 2015, pp. 5783–5790. 589
590
591
33. Sier, H.; Li, Q.; Yu, X.; Peña Queralta, J.; Zou, Z.; Westerlund, T. A benchmark for multi-modal lidar slam with ground truth in gnss-denied environments. *Remote Sensing* **2023**, *15*, 3314. 592
593
34. Yang, Y.; Xu, B.; Li, Y.; Schwertfeger, S. The slam hive benchmarking suite. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 11257–11263. 594
595

35. Furgale, P.; Rehder, J.; Siegwart, R. Unified temporal and spatial calibration for multi-sensor systems. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2013, pp. 1280–1286. 596
597
36. Cui, J.; Schwertfeger, S. CP+: Camera Poses Augmentation with Large-scale LiDAR Maps. In Proceedings of the 2022 IEEE 598
International Conference on Real-time Computing and Robotics (RCAR). IEEE, 2022, pp. 69–74. 599
37. Ouster, I. ouster OS0-128 REV6 datasheet. <https://data.ouster.io/downloads/datasheets/datasheet-rev06-v2p5-os0.pdf>, 2024. [Accessed 27-03-2024]. 600
601
38. insta360. Insta360 ONE RS 1-Inch 360 Edition – Dual 1-Inch Sensor 6K 360 Camera — insta360.com. <https://www.insta360.com/product/insta360-ones/1inch-360>, 2024. [Accessed 27-03-2024]. 602
603
39. insta360. Insta360 Titan — insta360.com. <https://www.insta360.com/product/insta360-titan>, 2024. [Accessed 27-03-2024]. 604
40. Zhao, X.; Yang, Z.; Schwertfeger, S. Mapping with reflection-detection and utilization of reflection in 3d lidar scans. In Proceedings 605
of the 2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). IEEE, 2020, pp. 27–33. 606
41. Zhao, X.; Schwertfeger, S. 3DRef: 3D Dataset and Benchmark for Reflection Detection in RGB and Lidar Data. In Proceedings 607
of the 2024 International Conference on 3D Vision (3DV), Los Alamitos, CA, USA, mar 2024; pp. 225–234. <https://doi.org/10.1109/3DV62453.2024.00009>. 608
609
42. Tancik, M.; Weber, E.; Ng, E.; Li, R.; Yi, B.; Kerr, J.; Wang, T.; Kristoffersen, A.; Austin, J.; Salahi, K.; et al. Nerfstudio: A Modular 610
Framework for Neural Radiance Field Development. In Proceedings of the ACM SIGGRAPH 2023 Conference Proceedings, 2023, 611
SIGGRAPH '23. 612
43. Li, Z.; Li, L.; Zhu, J. Read: Large-scale neural scene rendering for autonomous driving. In Proceedings of the Proceedings of the 613
AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 1522–1529. 614
44. Cui, J.; Cao, J.; Zhong, Y.; Wang, L.; Zhao, F.; Wang, P.; Chen, Y.; He, Z.; Xu, L.; Shi, Y.; et al. LetsGo: Large-Scale Garage Modeling 615
and Rendering via LiDAR-Assisted Gaussian Primitives. *arXiv preprint arXiv:2404.09748* **2024**. 616
45. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE 617
transactions on image processing* **2004**, *13*, 600–612. 618
46. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. 619
In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595. 620

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual 621
author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to 622
people or property resulting from any ideas, methods, instructions or products referred to in the content. 623