Comparison of State-of-the-Art SfM 3D Reconstruction Methods

Sean Hiroki Flynn Dr. Huaijin (George) Chen

M.S. Project Final Report
Department of Information and Computer Sciences, University of Hawai'i at Mānoa

ABSTRACT

3D reconstruction from 2D images is a fundamental problem in computer vision, with applications in robotics, augmented and virtual reality, and spatial computing. Recent AI-based methods have changed how reconstruction is done, but there are not many direct comparisons between these new methods and the classical Structure-from-Motion (SfM) pipelines that have been used for years. This project compares three recent AI methods (VGGT, Pi3, and Depth Anything 3) against COLMAP, a widely used classical SfM pipeline. The dataset is a custom set of four indoor scenes captured with an iPhone 15 Pro. Depth point clouds from a Gemini 335Lg stereo camera are used as ground truth. After scale alignment and ICP registration in CloudCompare, the reconstructions are evaluated using Chamfer Distance, Hausdorff Distance, accuracy, completeness, F-score, and reconstruction time. The results show that no single method dominates. COLMAP gets the best overall F-score because it produces more complete reconstructions, while Depth Anything 3 has the highest accuracy but lower completeness. The AI methods are clearly faster and handle low-texture scenes better, but they are sensitive to scene type. Overall, the comparison shows that classical and AI methods have different strengths, and choosing one depends on what the scene looks like and how much time you have.

1. INTRODUCTION

3D reconstruction is the task of figuring out the three-dimensional scene geometry from 2D images. Accurate 3D models are crucial for applications such as robotics, autonomous vehicles, and digital twins of buildings. Even though researchers have been working on this for several years, getting high-quality, dense reconstructions from a small number of photos is still hard.

The classical approach finds matching keypoints across images, figures out the camera position for each shot, and then triangulates the 3D structure using geometry. COLMAP gives accurate results when you have enough textured images, but it fails on smooth or texture-less surfaces. The newer approaches such as VGGT, Pi3, and Depth Anything 3, use neural networks that have been trained on large amounts of 3D data. These models take a few images and predict the 3D geometry directly, without computing the math. They are fast and they handle hard cases better. The geometry is essentially a learned guess, and they do not always behave the way you would expect.

It is not obvious when you should use which approach, so this project runs both side by side on the same scenes. Four pipelines (COLMAP, VGGT, Pi3, and Depth Anything 3) were tested on a custom dataset, with stereo depth as the reference. The goal is not to determine the best method. It is to figure out where each method works well, where it fails, and the trade-offs when selecting one for real-world applications.

2. METHODOLOGY

2.1 Dataset Collection

Five indoor scenes were captured for the analysis: an ukulele (day/night), an artificial plant (day), an electric fan (day), a lab chair, and a mannequin head. Each one was recorded as a 360 degree walk-around video with an iPhone 15 Pro at 1080p / 30 frames per second. The mannequin head was left out of the quantitative analysis because COLMAP could not reconstruct it. The failure is actually a useful result on its own and it will be mentioned in the Discussion. Frames were pulled from each video using ffmpeg. The AI pipelines used about 64 to 128 frames per scene, while COLMAP used about 80 to 300 frames so that its feature matching had enough overlap to work with.

2.2 Ground Truth

For ground truth, Orbbec Gemini 335Lg stereo camera was connected to a Windows 10 laptop and recorded the depth stream for each scene as a ROS1 `.bag` file using the Orbbec Viewer. A Python script was written (`extract_images.py`) to pull out the depth frames as 16-bit PNGs and log basic stats like depth min, max, and number of non-zero pixels. The depth images were back-projected into a 3D point cloud, and the point cloud is what each method gets compared against [5].

2.3 Reconstruction Pipelines

VGGT (Visual Geometry Grounded Transformer). VGGT is a feed-forward neural network that directly predicts all the key 3D attributes of a scene, including camera parameters, point maps, depth maps, and 3D point tracks, from one, a few, or hundreds of input views. It was released by Meta and Oxford's Visual Geometry Group in 2025. Unlike traditional SfM, VGGT skips the intermediate optimization steps and runs everything in a single forward pass through a 1.2 billion parameter transformer. The main appeal is speed. It can reconstruct a scene in under a second while still matching or beating methods that rely on geometric post-processing [1].

Pi3 (π^3). Pi3 is a feed-forward neural network for visual geometry reconstruction that breaks the reliance on a fixed reference view that earlier methods like DUSt3R and VGGT depend on. It uses a fully permutation-equivariant architecture to predict affine-invariant camera poses and scale-invariant local point maps without any reference frames. This means that the model is robust to the order of the input images, which removes a common failure mode where a poorly chosen reference view causes the whole reconstruction to break down [2].

Depth Anything 3. Depth Anything 3 (DA3) is a model that predicts spatially consistent geometry from any number of visual inputs, with or without known camera poses. It uses a DINOv2 encoder as the backbone and a unified depth-ray prediction target instead of separate task-specific heads. Earlier versions of Depth Anything were focused on monocular depth estimation, but DA3 extends the framework to handle single images, stereo pairs, multi-view collections, and video as a single foundation model. It was released by ByteDance Seed in late 2025 [3].

COLMAP. COLMAP is a classical Structure-from-Motion and Multi-View Stereo pipeline released in 2016 by Schönberger and Frahm. It extracts Scale-Invariant Feature Transform (SIFT) keypoints from each input image, finds matches between image pairs, incrementally building a sparse 3D point cloud and camera poses through triangulation and bundle adjustment, and then runs a Multi-View Stereo (MVS) step (Patch-Match Stereo and Stereo Fusion) to produce a dense point cloud. It has been the standard SfM tool in research for years and is the baseline the AI methods are compared against [4].

All the AI reconstructions were run on the Koa HPC cluster through Jupyter Lab on an NVIDIA RTX A4000 GPU. VGGT, Pi3, and Depth Anything 3 were run on the extracted frames for each scene, runtime was recorded, then GLB meshes were downloaded. COLMAP needs more memory, so an NVIDIA L40 (48 GB VRAM) was used. The full COLMAP pipeline: feature extraction, exhaustive matching, sparse reconstruction, im-

age undistortion, Patch Match Stereo, and Stereo Fusion was used. The output was a dense PLY point cloud and a runtime measurement for each scene.

2.4 Format Conversion and Alignment

The four pipelines all output reconstructions in different formats and coordinate systems, so it was necessary to convert and align everything prior to comparisons. The GLB meshes from the AI methods were converted to PLY using a web-based 3D model converter. After conversion, each reconstruction and its corresponding stereo ground truth was loaded into CloudCompare, scaled to match the ground truth, and then manually registered using Iterative Closest Point (ICP).

2.5 Evaluation Metrics

Once everything was aligned, each reconstruction to its ground truth was compared using the standard set of metrics: Chamfer Distance, Hausdorff Distance, accuracy, completeness, and F-score. They are defined below.

Chamfer Distance is the average distance between the points in the reconstruction and the points in the ground truth, lower is better [6]. Hausdorff Distance is the largest distance between the reconstruction and the ground truth, lower is better [7]. Accuracy asks: of all the points that the model predicted, what percentage are close to the ground truth? Higher is better [8]. Completeness asks: of all the ground truth points, how much of the scene did the method cover? Higher is better [8]. F-score combines accuracy and completeness into one number. The method needs to perform well on both to score high on this metric, so it is the standard overall quality metric [9].

Reconstruction time was also tracked as a separate measure of how practical each method is. All the metrics were computed by an `evaluate_metrics.py` script. This loads the aligned PLY files and prints a summary for each scene.

3. RESULTS

Table 1 shows the quantitative benchmark results across all methods and scenes. A few patterns we can see. COLMAP has the best completeness and the best overall F-score, which makes sense because it has more frames to work with and its dense MVS step fills in a lot of surface area. Depth Anything 3 has the lowest Chamfer and Hausdorff distances and the highest accuracy, meaning the points it predicts are close to the ground truth, but it does not predict as many points, so its completeness is the lowest of the four. The AI methods are also much faster than COLMAP. Depth Anything 3 finishes in under 80 seconds per scene, while COLMAP takes around 30 minutes.

4. QUALITATIVE RESULTS

CloudCompare was used for both ICP alignment and visual inspection. Figure 1 shows side-by-side renderings of the input image versus the 3D reconstruction. Each method has its own limitations:

- COLMAP outputs the most complete surfaces, but it gets noisy on low-texture areas like the back of the lab chair.
- VGGT and Pi3 produce clean geometry from fewer frames, but the scale sometimes drifts and they occasionally hallucinate backgrounds that are not really there.
- Depth Anything 3 puts points in the right places but leaves big gaps on the sides and back of objects due to its monocular nature, which is why its completeness score is so low.
- On the textureless mannequin head, COLMAP could not generate any reconstruction, while all three AI methods worked fine. This case captures the main trade-off in the whole project well, and it will be discussed more in depth later.

Table 1: Quantitative benchmark results across all methods and scenes.

Method	**Scene**	**Chamfer ↓** (mm)	**Hausdorff ↓** (mm)	**Accuracy ↑**	**Completeness ↑**	**F-Score ↑**	**Reconstruction Time (s) ↓**
VGGT	Ukulele Day	10.79	1163.56	88.39%	65.93%	75.53%	19.19 s
Pi3	Ukulele Day	5.18	423.73	91.42%	62.23%	74.05%	35.26 s
DepthAnything3	Ukulele Day	10.86	268.15	88.33%	66.53%	75.90%	2.25 s
COLMAP	Ukulele Day	5.52	464.69	91.34%	71.2%	80.02%	24.802 min
VGGT	Ukulele Night	7.50	1314.51	91.17%	71.51%	80.15%	15.36 s
Pi3	Ukulele Night	23.02	1216.95	93.01%	69.49%	79.55%	23.42 s
DepthAnything3	Ukulele Night	0.17	207.94	98.69%	45.04%	61.85%	1.27 s
COLMAP	Ukulele Night	3.27	658.20	87.38%	80.01%	83.54%	24.387 min
VGGT	Artificial Plant Day	10.00	765.90	77.4%	46.61%	58.18%	13.85 s
Pi3	Artificial Plant Day	13.05	688.79	78.16%	49.75%	60.80%	21.02 s
DepthAnything3	Artificial Plant Day	0.72	132.17	98.85%	46.63%	63.37%	2.99 s
COLMAP	Artificial Plant Day	9.28	572.12	78.23%	60.8%	68.42%	21.593 min
VGGT	Electric Fan Day	2.11	1485.46	86.66%	82.11%	84.32%	20.66 s
Pi3	Electric Fan Day	6.08	1678.49	81.35%	91.46%	86.11%	34.88 s
DepthAnything3	Electric Fan Day	1.20	228.92	95.98%	39.99%	56.46%	3.42 s
COLMAP	Electric Fan Day	8.50	1277.66	80.08%	84.59%	82.27%	31.008 min
VGGT	Lab Chair	8.99	1554.96	41.69%	34.81%	37.94%	19.77 s
Pi3	Lab Chair	12.81	1116.10	23.99%	23.35%	23.66%	31.69 s
DepthAnything3	Lab Chair	19.92	534.96	64.23%	36.42%	46.48%	5.17 s
COLMAP	Lab Chair	13.12	943.01	58.72%	64.19%	61.34%	29.509 min

↓ lower is better ↑ higher is better All distances in millimeters Mannequin Head: qualitative only

5. DISCUSSION

The mannequin head case is arguably the most interesting result of the whole project. COLMAP relies on SIFT keypoints and geometric verification between image pairs, and on a smooth, texture-less surface there just are not enough repeatable features to find. So, the pipeline ends. The AI methods, which were trained on millions of 3D scenes, generate reasonable geometry from the same input video. This is exactly the kind of situation where the AI methods' training data excels. They have seen so many objects before that they can make a good guess at the geometry even without clear features to match. This is a good example of why the AI methods are not just a faster version of COLMAP. They are solving the problem in a completely different way.

The combined metrics support this. COLMAP outperforms on completeness and F-score because it had enough frames to fully use its dense Multi-View Stereo stage, and because the textured scenes (Ukulele, Artificial Plant, Electric Fan, Lab Chair) gave it lots of features to match. Depth Anything 3 wins on accuracy but loses on completeness because monocular depth predictions can be sharp where the camera looks but cannot recover surfaces the model never sees from a confident angle. VGGT and Pi3 lie in between: not the most accurate, not the most complete, but a good balance and around 10× faster than COLMAP. In practice, which method you would select depends on the task. COLMAP for high-quality scans of textured scenes when there is plenty of input. AI methods for fast reconstruction, sparse inputs, or low-texture surfaces.

The frame count difference is also worth noting. The AI pipelines worked well with 64 to 128 frames, while COLMAP needed 80 to 300 to get a similar density. This makes sense given how each one works. The quality of COLMAP improves the more matching features it can find between images, so more frames mean more matches and better geometry. The AI models do not need that many frames because they already recognize what objects tend to look like in 3D. Once they have enough views to make a confident prediction, adding more does not help much.

6. LIMITATIONS

The dataset is small and limited to indoor scenes. The Gemini 335Lg's ideal depth range is 0.25 to 6 meters, which sits outside the working distances necessary for outdoor capture. More importantly, its active IR projector is overwhelmed by ambient sunlight, which is a known issue for active stereo cameras in outdoor lighting. This means the camera cannot generate reliable ground truth in the conditions where outdoor scenes would be captured.

Low light was another issue. The stereo camera performed poorly on the Ukulele Night scene, where the captured depth map did not resemble the silhouette of a ukulele. All four pipelines were run with default or recommended settings. Tuning would almost certainly shift the relative numbers, particularly for COLMAP, where dense reconstruction is sensitive to matcher and stereo parameters.

7. CONCLUSION

Across four indoor scenes, no single method dominates across all metrics. COLMAP had the best completeness and overall F-score. Depth Anything 3 had the most accurate per-point geometry. VGGT and Pi3 had the best balance of speed and quality. The texture-less mannequin head exposed COLMAP's reliance on classical features in a way that no combined metric could. In conclusion, AI methods and classical methods do not really replace each other. They complement each other.

8. FUTURE WORK

For future work, the dataset should be extended to outdoor scenes by using a depth sensor better suited for outdoor capture, such as the Stereolabs ZED 2i or an Intel RealSense D455, which are both designed to handle bright sunlight and longer working distances. More texture should be added to the Mannequin Head as well as objects with a wider range of materials.

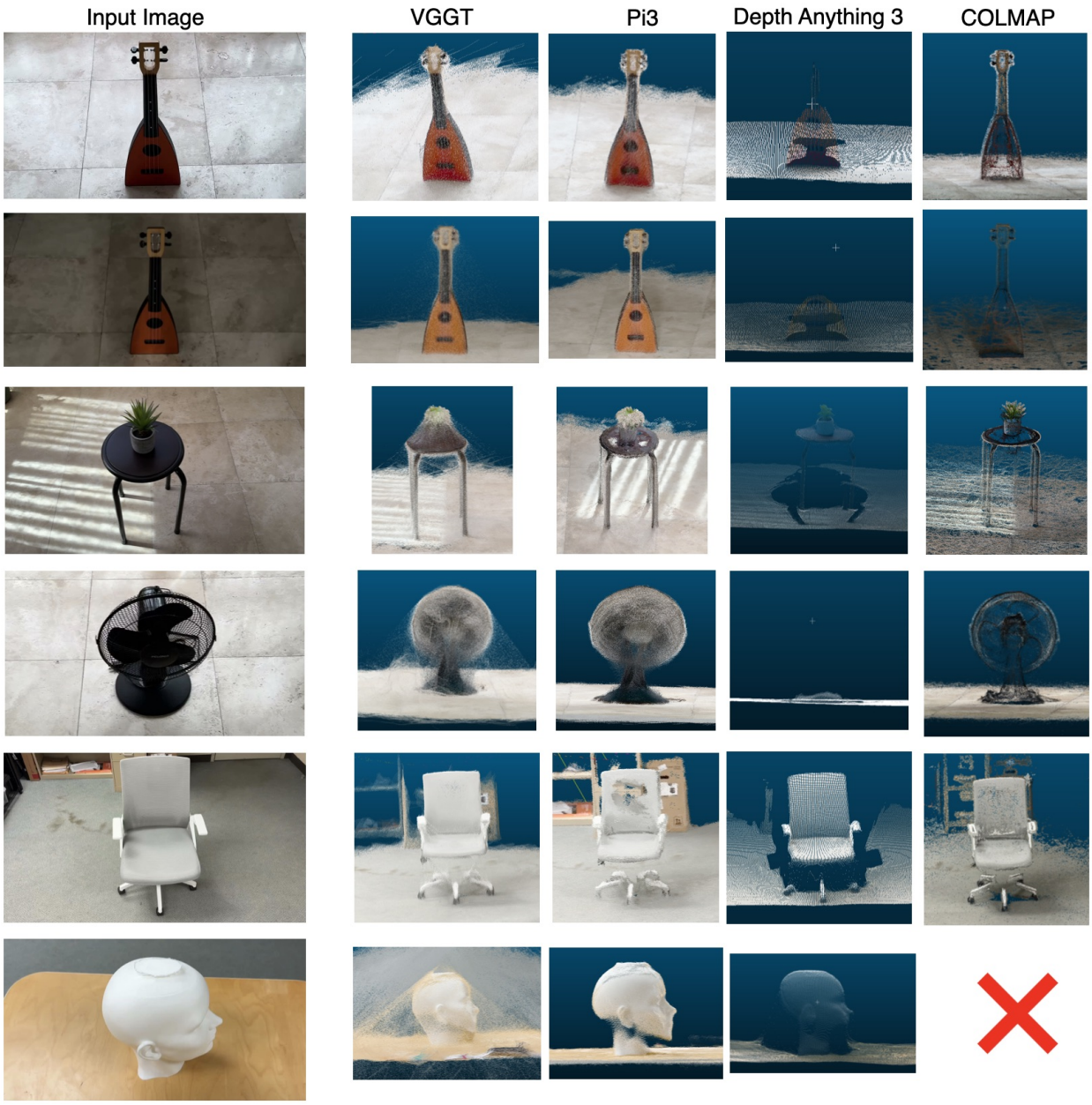

Figure 1: CloudCompare visualizations of input image vs. 3D reconstruction

REFERENCES

[1] Wang, Jianyuan, et al. "VGGT: Visual Geometry Grounded Transformer." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[2] Wang, Yifan, et al. "π^3: Scalable Permutation-Equivariant Visual Geometry Learning." *arXiv preprint arXiv:2507.13347*, 2025.

[3] Lin, Haotong, et al. "Depth Anything 3: Recovering the Visual Space from Any Views." *arXiv preprint arXiv:2511.10647*,v2025.

[4] Schönberger, Johannes Lutz, and Jan-Michael Frahm. "Structure-from-Motion Revisited." *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Gemini 335Lg Stereo Camera, Orbbec Inc.

[6] P. Alonso, T. Li, and C. Li, "GeoCD: A Differential Local Approximation for Geodesic Chamfer Distance," *arxiv.org*, 2025. https://arxiv.org/html/2506.23478v1 (accessed May 10, 2026).

[7] "Hausdorff Distance - an overview | ScienceDirect Topics," *www.sciencedirect.com*. https://www.sciencedirect.com/topics/computer-science/hausdorff-distance

[8] I. Petrovska and B. Jutzi, "Vision through Obstacles—3D Geometric Reconstruction and Evaluation of Neural Radiance Fields (NeRFs)," *Remote Sensing*, vol. 16, no. 7, pp. 1188–1188, Mar. 2024, doi: https://doi.org/10.3390/rs16071188.

[9] Z. Leng, T. Birdal, X. Liang, and F. Tombari, "HyperSDFusion: Bridging Hierarchical Structures in Language and Geometry for Enhanced 3D Text2Shape Generation," 2024. Accessed: May 10, 2026. [Online].