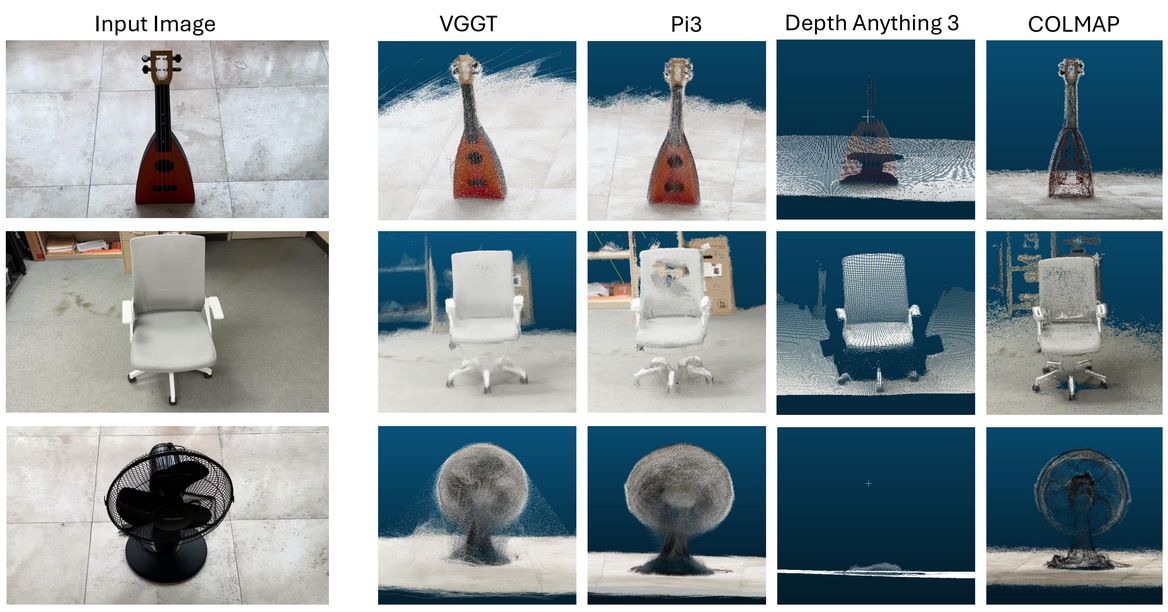
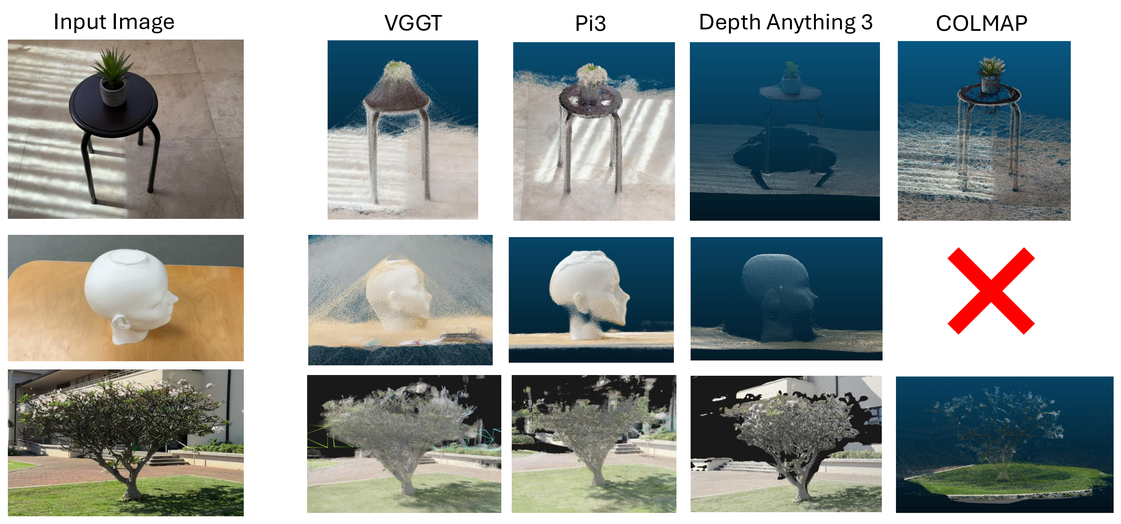
Comparison of State-of-the-Art SfM 3D Reconstruction Methods

Sean Hiroki Flynn, MS
Dr. Huaijin (George) Chen
University of Hawai'i at Mānoa, Department of Information and Computer Sciences





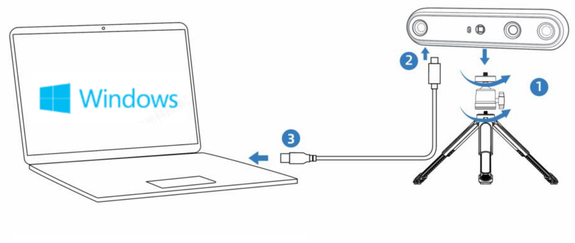
Introduction

3D reconstruction from 2D images is a fundamental problem in computer vision with applications in robotics, AR/VR, and spatial computing.
Recent AI-based methods challenge classical pipelines, but rigorous comparisons under controlled conditions are limited.
This project benchmarks VGGT, Pi3, Depth Anything 3, and COLMAP on a custom indoor scene dataset with stereo depth data as ground truth.

VGGT — transformer-based feed-forward 3D reconstruction
Pi3 — AI-based multi-view reconstruction
Depth Anything 3 — monocular depth estimation extended to 3D
COLMAP — classical SfM with multi-view stereo dense reconstruction

Methods

iPhone 15 Pro – 1080p 30fps
(Scene capture)



ORBBEC
Viewer

Gemini 335Lg
stereo vision camera
(Depth data capture)

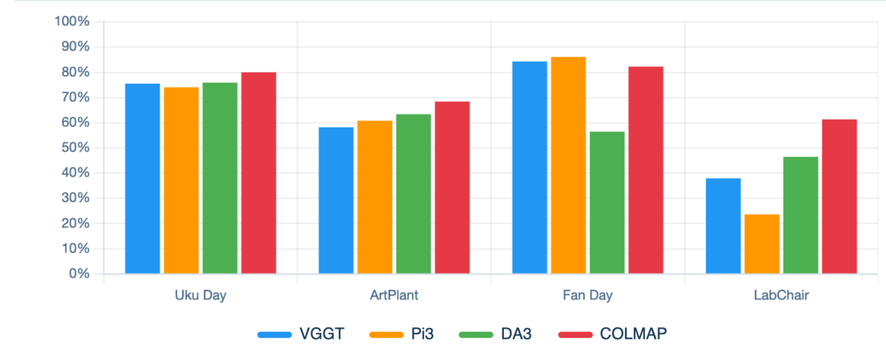
5 INDOOR SCENES (frames extracted using ffmpeg at 3 fps ≈ 90 frames/scene)

1. Ukulele — My House
2. Electric Fan — My House
5. Mannequin Head — CIRP Lab at POST
3. Artificial Plant — My House
4. Lab Chair — CIRP Lab at POST
\+
Depth data — Stereo PLY

Evaluation Pipeline

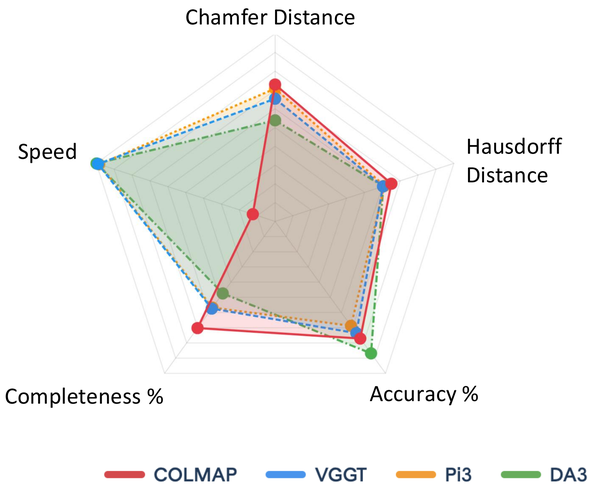
iPhone video → Frame extraction (ffmpeg / 3fps)
→ VGGT | Pi3 | Depth Anything 3 | COLMAP
→ CloudCompare (Scale & ICP Alignment) → evaluate_metrics.py

Results



F-Score % by Scene

↑ Higher is better



Metric Definitions:

Chamfer – Avg distance from point-to-point vs ground truth

Hausdorff – Max distance from point-to-point vs ground truth

Accuracy – How close predicted points are to ground truth

Completeness – How much of the scene was captured?

F-Score – Combined measure of Accuracy & Completeness

Speed – Total reconstruction time

Challenges

VGGT

Pi3

Why Does This Matter?

- Autonomous vehicles use 3D reconstruction to map environments in real time
- Surgical robotics and medical imaging use 3D models for precision navigation
- AR/VR and spatial computing require fast, accurate reconstruction to feel real
- Disaster response drones use 3D mapping to assess structural damage

Conclusions

- No single method dominates across all scenes and metrics
- COLMAP has best overall F-score - most complete reconstructions
- Texture-less surfaces (Mannequin Head) break COLMAP entirely
  - SIFT requires distinct keypoints
- LabChair is clearly the hardest scene for all methods, likely due to complex geometry and obstructions
- Depth Anything 3 achieves high accuracy but low completeness
- Future work: evaluate on larger outdoor datasets, include NeRF-based methods

Contact

Sean Hiroki Flynn
University of Hawai'i at Mānoa
Email: sflynn7@hawaii.edu
Website: https://github.com/sflynn7

References

1. Wang, Jianyuan, et al. "VGGT: Visual Geometry Grounded Transformer." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.Pi3: Fast and Scalable Point Cloud Reconstruction from Images. 2025.
2. Lin, Haotong, et al. "Depth Anything 3: Recovering the Visual Space from Any Views." arXiv preprint arXiv:2511.10647, 2025.
3. Wang, Yifan, et al. "π³: Scalable Permutation-Equivariant Visual Geometry Learning." arXiv preprint arXiv:2507.13347, 2025.
4. Schönberger, Johannes Lutz, and Jan-Michael Frahm. "Structure-from-Motion Revisited." Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
5. Gemini 335L Stereo Camera, Orbbec Inc.