3D SCENE RECONSTRUCTION USING SATELLITE IMAGES

Sumit Raut
Department of Computer Science
University of California, Irvine
Irvine, CA 92697, USA
scraut@uci.edu

ABSTRACT

This project aim to reconstruct 3D satellite imagery using Structure-from-Motion (SfM) and Multi-view stereo techniques. We collected multi-temporal and 3D-enhanced satellite images of the UCI campus from Google Earth Pro and processed them using COLMAP to generate sparse and dense point clouds. Evaluation across three datasets showed that reconstructions benefit from topographic detail, with 3D enabled topographic imagery achieving the lowest reprojection error. Challenges included limited parallax and the lack of native support for learned depth fusion in COLMAP. Despite these constraints, our results demonstrate that 3D scene reconstruction is feasible using SfM when supported by richer 3D imagery or additional ground-based views. Future work can explore enhancements such as semantic masking, more flexible dense fusion pipelines, and alternative approaches that incorporate externally fused depth maps for improved reconstruction quality.

1 OVERVIEW OF 3D SCENE RECONSTRUCTION

1.1 DEFINITION

3D reconstruction refers to the process of generating a digital three-dimensional representation of a physical scene or object from two-dimensional images. This transformation typically relies on estimating depth information either through geometric triangulation or via learned priors. In the context of satellite imagery, 3D reconstruction aims to extract structural features—such as buildings and terrain elevations—using images captured from different viewpoints over time. Applications include urban planning, environmental monitoring, disaster response, and digital preservation of geographic data. Despite the promise, satellite imagery poses unique challenges due to limited parallax, atmospheric distortion, and frequent occlusions, making robust reconstruction a non-trivial task.

1.2 RELATED WORKS

Traditional approaches to 3D reconstruction rely on multi-view geometry, with Structure-from-Motion (SfM) and Multi-View Stereo (MVS) forming the foundation of most classical pipelines. Tools like COLMAP automate these steps through feature detection (e.g., SIFT), exhaustive pairwise matching, and global bundle adjustment Schönberger & Frahm (2016). However, such pipelines are sensitive to viewpoint variation and perform poorly when applied to satellite imagery due to its limited parallax and top-down perspective.

To overcome these challenges, several methods have been proposed. LuxCarta's CREAS-MAP demonstrates large-scale 3D building model generation from multi-angle satellite views, combining geometric processing with cartographic modeling tools LuxCarta. Similarly, Facciolo et al. introduced a robust stereo pipeline for multi-date satellite imagery, exploiting temporal diversity to simulate wide-baseline stereo pairs de Franchis et al. (2020). Their work, along with that of Bullinger et al. Bullinger et al. (2021), shows that satellite imagery spanning different dates can provide enough parallax for accurate surface reconstruction.

In parallel, deep learning-based monocular depth estimation has gained popularity for scenarios where multi-view consistency is lacking. Models such as MiDaS Ranftl et al. (2021), ZoeDepth Bhat et al. (2023), and DPT Ranftl et al. (2022) predict per-image depth maps using learned visual priors and have been applied to indoor, outdoor, and remote sensing scenes. Hybrid approaches are also emerging. The VisSat project, for instance, adapts traditional vision reconstruction pipelines to satellite domains Zhang et al. (2019), while more advanced work proposes virtual Rational Polynomial Coefficient (RPC) modeling using homography-based georeferencing for aligning stereo pairs across views Nakamura et al. (2023).

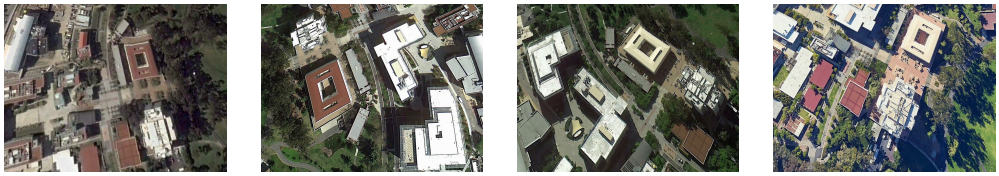
These efforts collectively suggest that combining classical geometry, learned depth priors, and temporal baselines enables more robust 3D reconstruction from satellite imagery, even under minimal parallax conditions.

2 Technical Approach

2.1 Data Collection

Data was captured using recordings of satellite images over the elevation of roughly 900ft using Google Earth Pro, focusing on the ICS Department area of the UCI campus. Frames were extracted(roughly every 60th frame in recording) to generate image dataset. To simulate multi-view geometry, recordings were taken while observing the scene along different directions—north, south, east, and west—ensuring sufficient visual overlap between frames. Three datasets were created to evaluate reconstruction quality under varying temporal and structural contexts:

1. **Temporal dataset**: Consisting of roughly 300 images collected from satellite archives spanning 2005–2025.



(a) 2005 (b) 2010 (c) 2017 (d) 2025

Figure 1: Temporal dataset samples from different years

2. **Current year dataset**: 50 high-resolution satellite images captured in 2025.

Figure 2: Satellite images captured in 2025

3. **3D-enhanced dataset**: Roughly 140 images from 2025 with terrain and 3D building enabled.

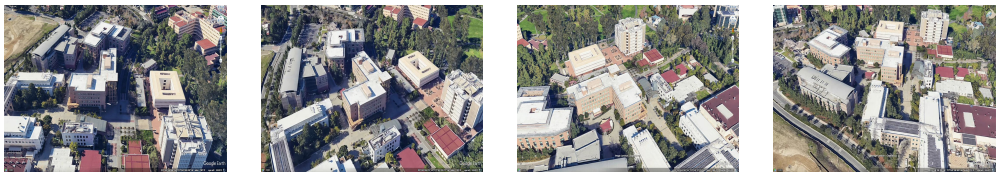


Figure 3: 3D building enabled images

These datasets provided diverse image baselines and enabled comparison of reconstruction performance under different imaging conditions.

2.2 Pre-Processing Satellite Images

To ensure compatibility with SfM pipelines and improve matching robustness, all satellite images were resized to a fixed resolution of 1024×768 using the Python Pillow library. Images were extracted at regular intervals (every 60th frame) from screen recordings in Google Earth Pro. This provided sufficient visual overlap and diversity. Datasets were then organized into separate colmap workspaces for different reconstruction experiments (temporal, current, 3D-enhanced).

2.3 Sparse Reconstruction

Sparse reconstruction involves generating a 3D point cloud and estimating camera poses using a set of overlapping 2D images. In our project, we used COLMAP—a widely-used SfM pipeline—to perform sparse reconstruction across all datasets: the temporal set, the current satellite view, and the 3D-enhanced imagery.

The SfM pipeline proceeds by first detecting keypoints in each image, extracting SIFT descriptors that are invariant to scale and rotation. These features are then exhaustively matched across all image pairs to find consistent correspondences. Once matches are established, the mapper incrementally registers cameras and triangulates the 3D positions of matching keypoints. A global bundle adjustment is then performed to refine camera intrinsics, poses, and 3D point positions by minimizing the reprojection error.

The result is a sparse point cloud that captures the geometric structure of the scene. Additionally, the intrinsic and extrinsic parameters of all successfully registered cameras are estimated. This sparse model returns initial point cloud for downstream tasks such as dense reconstruction or depth map alignment.

Technical Steps:

- **Feature Extraction:** COLMAP extracted SIFT features from all images. These descriptors capture local structure and are robust to changes in scale and illumination.
- **Feature Matching:** An exhaustive matcher compared all image pairs to identify shared feature points. These matches are critical for geometric triangulation and are stored in the COLMAP database.
- **Sparse Mapping:** COLMAP's incremental mapper registered images one by one. Using RANSAC and epipolar geometry constraints, camera poses were estimated, and 3D landmarks were triangulated from matched features.
- **Bundle Adjustment:** A nonlinear optimization step jointly refined the 3D structure and camera parameters to minimize reprojection error. This improved both the fidelity of 3D points and the accuracy of the camera extrinsics.
- **Model Export:** The resulting sparse model, stored in binary format, was optionally converted to `.PLY` for visualization using external tools. The model includes files for cameras, images, and 3D points.

2.4 Dense Reconstruction

Dense reconstruction aims to produce a detailed 3D model by estimating per-pixel depth from multiple calibrated images. In this project, we leveraged COLMAP's multi-view stereo (MVS) pipeline to densify the sparse point cloud generated during the Structure-from-Motion phase. Dense reconstruction enhances scene fidelity by producing a much higher resolution point cloud that captures finer geometric details.

COLMAP's dense pipeline consists of several stages: image undistortion, depth estimation using PatchMatch Stereo, and depth fusion to generate a dense 3D point cloud. Due to hardware limitations (absence of CUDA support), the full pipeline could not be completed using COLMAP alone

on some systems. Therefore, execution was done on local system and Google colab for using CUDA environment.

Pipeline:

- **Image Undistortion:** Input images were undistorted using the calibrated intrinsic parameters obtained during the sparse reconstruction. This step prepares the images for stereo matching by removing lens distortion and rectifying image geometry.
- **PatchMatch Stereo:** COLMAP applies a multi-view PatchMatch algorithm to compute dense depth maps for each undistorted image. This method estimates depth by comparing matching windows across multiple views using a randomized optimization strategy. Geometric consistency checks are used to suppress outliers and improve robustness. This stage requires CUDA and could not be executed on all machines.
- **Stereo Fusion:** The individual depth maps from PatchMatch were fused into a single global dense point cloud. The fusion algorithm merges consistent depth estimates from different views, weighted by confidence and visibility. The resulting 3D point cloud contains significantly more points than the sparse model, capturing surface geometry at higher resolution.
- **Surface Meshing:** After depth fusion, COLMAP supports converting the dense point cloud into a mesh using Poisson or Delaunay triangulation. This produces a watertight surface model useful for rendering or simulation purposes.
- **Alternative: Monocular Depth Back-Projection:** To address hardware limitations, we also employed MiDaS (DPT Hybrid model) to predict monocular depth maps for each input image. These depth maps were then back-projected into 3D using the intrinsic matrix K and camera-to-world transformation $[R|t]$ provided by COLMAP. The resulting 3D points from each image were merged into a global point cloud.
- **Visualization and Export:** All dense point clouds were visualized using Open3D,MeshLab and COLMAP.

3 RESULTS

3.1 SPARSE RECONSTRUCTION EVALUATION

The performance of the sparse reconstruction pipeline was evaluated using three key metrics: the number of registered images, the number of triangulated 3D points, and the mean reprojection error. These values reflect the geometric consistency and completeness of the reconstructed scene. All evaluations were conducted using COLMAP's internal reconstruction statistics and exported model files.

Evaluation Metrics:

- **Registered Images:** This metric indicates how many input images were successfully localized in 3D space. A high number implies robust feature matching and good overlap between views.
- **3D Points:** The total number of successfully triangulated points reflects the structural density of the scene captured through multi-view correspondence. More points generally indicate higher coverage of the scene geometry.
- **Mean Reprojection Error:** This is the average distance (in pixels) between the observed image keypoints and the projected 3D points back onto the images using estimated camera poses. Lower error suggests better geometric accuracy and a well-optimized bundle adjustment.

Results: The table below summarizes the quantitative results for each dataset:

Dataset Type	Registered Images	3D Points	Mean Reproj. Error
3D-enabled Satellite Images	132	112,177	0.7343
Temporal Satellite Images (2005–2025)	279	309,103	1.0924
Current Satellite Images (2D only)	47	88,436	1.1614

Table 1: Sparse reconstruction statistics across 3 pipelines.

Analysis: As shown in Table 1, the temporal dataset achieved the highest number of registered images and triangulated 3D points. This can be attributed to the increased visual diversity provided by multi-year image captures, which improved feature matching across wider baselines. However, the 3D-enabled dataset achieved the lowest mean reprojection error, likely due to the inclusion of topographic details and artificial building extrusion, which improved spatial consistency.

In contrast, the current (2025) dataset had the fewest registered images and the highest reprojection error. This was expected due to the limited number of views and minimal parallax in nadir-only captures, which restricts COLMAP's ability to triangulate accurate 3D structure.

3.2 Dense Reconstruction Evaluation

To evaluate the quality and completeness of the dense 3D reconstructions, we analyzed the final fused point clouds generated for each pipeline. The evaluation focused on three key metrics:

- **Number of Points:** Total number of points in the dense point cloud after fusion, reflecting the reconstruction density.
- **Axis-Aligned Bounding Box (AABB):** The spatial extent of the point cloud in 3D space, used to compute overall scene coverage and volume.
- **Volume:** The volume of the AABB, indicating how much space the reconstructed geometry occupies.

These metrics were computed using the `Open3D` library. The bounding box dimensions also provide qualitative insight into the scale, spread, and compactness of the reconstructed scene.

Pipeline	Points	Volume (m^3)	BBox Min	BBox Max
Current Year Satellite (30 images)	572,388	973.61	[-7.50, -1.58, 4.95]	[3.85, 7.56, 14.33]
Temporal Satellite (2005–2025, 300 images)	3,407,540	18,512.80	[-6.06, -2.93, 0.33]	[6.04, 12.68, 98.36]
3D-enabled Satellite (115 images)	166,123	1,497.63	[-5.95, -2.39, -0.87]	[8.49, 10.00, 7.50]

Table 2: Dense reconstruction statistics: point count, bounding box, and spatial volume

Analysis: The temporal dataset yielded the highest number of 3D points and the largest reconstruction volume. However qualitative analysis using images in later stage shows it did not capture structures that well and rather gave random large 3D volume.

In contrast, the 3D-enabled dataset produced fewer points but achieved a higher geometric compactness, reflected in a smaller yet denser bounding volume. This dataset also displayed cleaner structure and sharper building outlines.

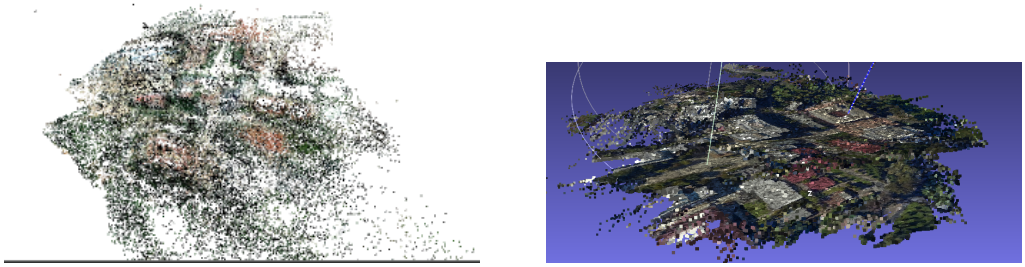
The current-year dataset, with only 30 images and no terrain detail, resulted in a moderate number of points and the smallest spatial extent. The scene mostly represented volume due to a flat surface generation with little depths visible for most of the structures.

Overall, these results confirm that 3D-enriched inputs significantly enhance dense reconstruction outcomes in satellite imagery settings.

3.3 Qualitative Evaluation

To complement the quantitative analysis, we present qualitative visualizations of the dense 3D reconstructions for each dataset. The following figures show two representative views per pipeline, rendered from the final fused point clouds. These visualizations reveal the structural integrity, density, and surface completeness of the reconstructions.

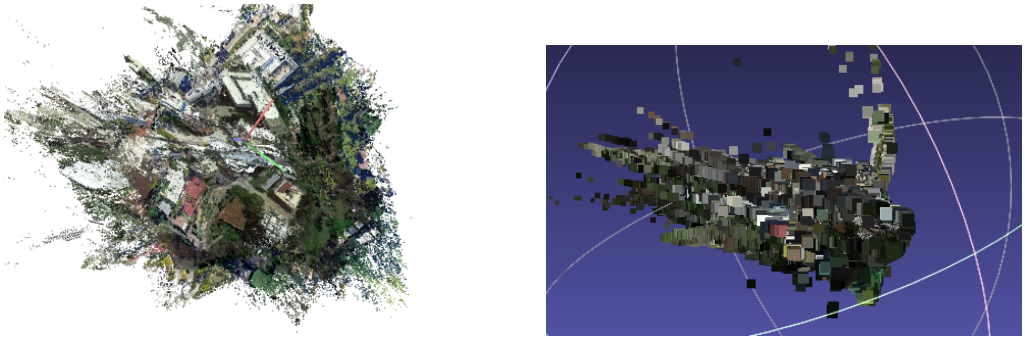
Current Year Satellite (50 images):



(a) COLMAP Visualization (b) MeshLab Visualization

Figure 4: Flat Surface type reconstruction observed from current year satellite imagery

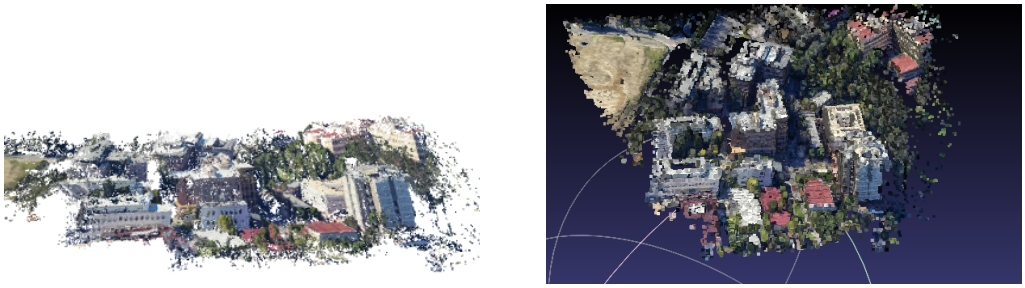
Temporal Satellite (2005–2025, 300 images):



(a) COLMAP Visualization (b) MeshLab Visualization

Figure 5: Sparse and noisy reconstruction due to inconsistent multi-year imagery

3D-enabled Satellite Imagery (140 images):



(a) COLMAP Visualization (b) MeshLab Visualization

Figure 6: Dense reconstruction from 3D-enabled imagery with best structural preservation

Irregular 3D reconstruction due to shadow in image:

(a) Original Image (b) 3D Reconstruction Patch

Figure 7: Dense reconstruction is not visible for areas having shadow in image.

Irregular 3D reconstruction trees near building

(a) Original Image (b) 3D Reconstruction Patch

Figure 8: Dense reconstruction not visible near walls of buildings close to trees.

Visual Observations: Based on these visualizations, the 3D-enabled dataset clearly produced the most geometrically consistent and visually complete reconstruction. In contrast, the temporal dataset suffered from severe misalignment and noise due to lighting variation and inconsistent viewpoints over time. The current dataset yielded high point density but was limited in vertical resolution. Also some aspects like shadow in images and objects near building affect the overall completness of 3D scene reconstruction.

4 Discussion

4.1 Challenges

This project faced several technical and domain-specific challenges during both sparse and dense reconstruction. A major limitation stemmed from COLMAP's lack of native support for integrating externally predicted depth maps—such as those generated by deep learning based image models—into its dense fusion pipeline. It does not offer an interface for directly incorporating learned depth priors, which limited our ability to build hybrid geometric-learning pipelines within its framework. Another significant challenge was the limited parallax inherent in satellite imagery. Since

most satellite images are captured from viewpoints with small angular differences, depth triangulation is difficult. As a result, reconstructions were often biased toward horizontal structures like rooftops and flat terrain, while vertical elements such as building walls were poorly captured or entirely missing. Temporal variability across datasets introduced additional complexity. For instance, images collected from different years often exhibited lighting changes, seasonal variation, and structural alterations. These inconsistencies degraded the effectiveness of feature matching and camera registration in the sparse pipeline.

4.2 Future Work

To improve the robustness and quality of future reconstructions, several enhancements like incorporating semantic segmentation and object masking into the reconstruction pipeline can be made. By using models such as Mask2Former or Segment Anything, it would be possible to isolate static structures while removing dynamic or irrelevant background elements such as trees, cars, and shadows. This could significantly improve the quality and interpretability of the resulting point clouds. Another feasible alternative for dense reconstruction is using frameworks such as OpenMVS or NeuralRecon, which support external depth map fusion and may offer more flexibility for integrating learned depth priors. These frameworks could be used in conjunction with COLMAP's sparse reconstructions, allowing a hybrid pipeline that combines geometric and learning-based cues more effectively.Addition of ground based images can also be tested to make up for shadowed areas in images for completeness of models. Together, these future directions highlight the potential of combining classical geometry with modern deep learning and satellite-specific priors to push the boundaries of 3D reconstruction in remote sensing contexts.

5 Conclusion

This project explored the feasibility and effectiveness of reconstructing 3D scenes from satellite imagery using a combination of classical Structure-from-Motion (SfM) techniques and Multi-view-stereo pipeline. Our findings indicate that traditional SfM methods can produce viable 3D reconstructions from satellite data, particularly when augmented with terrain and structural detail, as in the 3D-enabled dataset. However, challenges such as limited parallax, poor vertical visibility, and the lack of integration between learned depth priors and geometric reconstruction pipelines restrict the overall completeness of the models. Despite these limitations, the experiments demonstrate that even moderately overlapping satellite images can be used to generate meaningful spatial representations of urban environments. Moving forward, integrating object-aware masking, adopting alternative dense fusion frameworks, and training depth models specifically for satellite imagery will be essential for advancing the state of 3D reconstruction.

References

Shariq Farooq Bhat et al. Zoedepth: Zero-shot transfer for monocular depth estimation. *arXiv preprint arXiv:2302.12288*, 2023.

S. Bullinger, C. Bodensteiner, and M. Arens. 3d surface reconstruction from multi-date satellite images. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLIII-B2-2021, pp. 313–320, 2021. doi: 10.5194/isprs-archives-XLIII-B2-2021-313-2021.

Carlo de Franchis, Enric Meinhardt-Llopis, and Gabriele Facciolo. Automatic 3d reconstruction from multi-date satellite images. *École Normale Supérieure Paris-Saclay*, 2020. Accessed: 2025-06-09.

LuxCarta. Creas-map: Urban modeling from satellite imagery. `https://www.luxcarta.com/blog/creas-map-3d-reconstruction`. Accessed: 2025-06-09.

Yuji Nakamura, Yuki Suga, and Kenichi Muraoka. 3d reconstruction from multi-view google earth satellite stereo images by generating virtual rpc based on 3d homography-based georeferencing. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLVIII-1/W2-2023, pp. 1075–082. Copernicus GmbH, 2023. doi: 10.

5194/isprs-archives-XLVIII-1-W2-2023-1075-2023. URL https://isprs-archives.copernicus.org/articles/XLVIII-1-W2-2023/1075/2023/. Accessed: 2025-06-09.

Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *arXiv preprint arXiv:2103.13413*, 2021.

Rene Ranftl et al. Dense prediction transformer. *arXiv preprint arXiv:2103.13413*, 2022.

Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. *CVPR*, 2016.

Kai Zhang, Jin Sun, and Noah Snavely. Leveraging Vision Reconstruction Pipelines for Satellite Imagery. In *ICCV Workshop on 3D Reconstruction in the Wild (3DRW)*, 2019.