Emory2MC: 3D Mesh Reconstruction and Voxel Rendering of Emory University in Minecraft

Thomas Ma[1] Alexander Liu[1] Akhil Arularasu[1]

[1]Emory University, Department of Computer Science

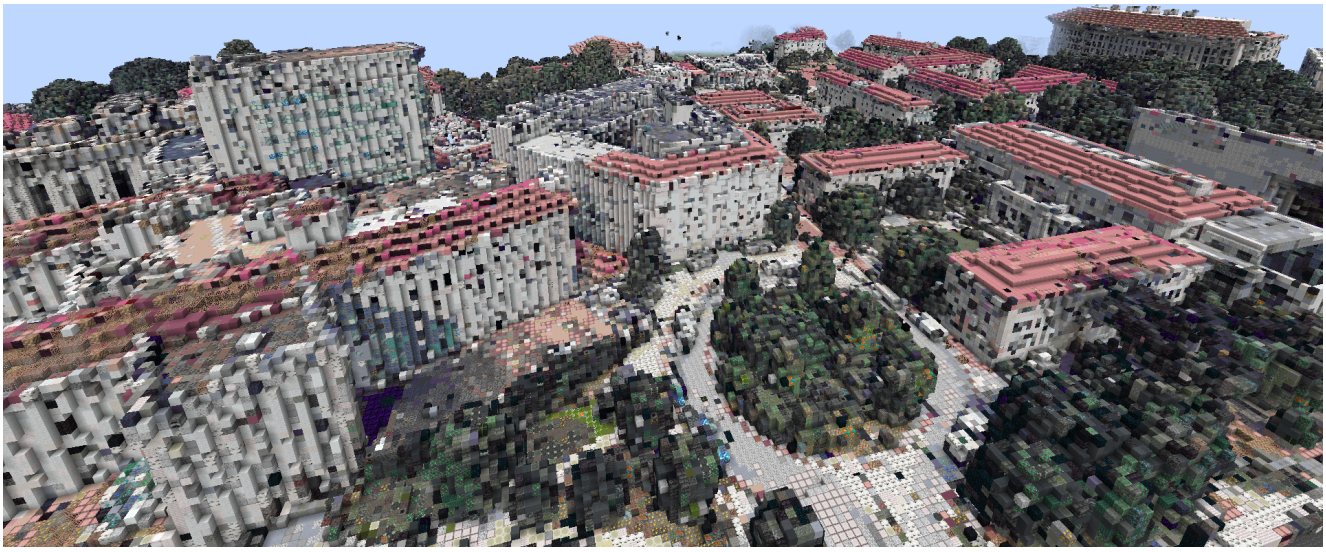

Figure 1. **Emory2MC** is an exploration of various 3D reconstruction approaches using video footage data of Emory University campus.

Abstract

In this paper, we seek to build an approach that is capable of representing Emory University in Minecraft. We explore several approach pipelines that constructs a 3D representation of an outdoor scene (Emory University), recovers a 3D mesh, and voxelizes and match Minecraft blocks by similarity. We then visually assess the capability of our pipelines before exploring next steps to improve our pipeline.

1. Introduction

Minecraft is arguably one of the most popular video games of all time, captivating a massive global audience with its procedurally generated, block-based, open-world environment. Its core mechanic revolves around the placement and breaking of blocks, fundamentally making it a three-dimensional grid structure. This structure is highly compatible with external tools and formats, notably allowing users to load complex pre-built structures through standardized file formats such as `.schem` (Schematic) and `.litematica`.

The possibility of representing real-world architecture—such as the campus of Emory University—within the **Minecraft** environment holds several compelling benefits. First, it offers a novel and highly engaging way to improve the institution's public perception and outreach, potentially fostering increased interest from prospective students in a medium they are already familiar with. Second, and more importantly from a technical standpoint, the successful conversion of real-world assets into a **Minecraft**-compatible 3D mesh is a powerful demonstration of the current state of video-to-3D reconstruction pipelines. Such pipelines have broad applications beyond gaming, including virtual tourism, urban planning, and digital preservation.

The challenge lies in accurately and efficiently generating a high-fidelity 3D mesh model from unstructured video footage. The field of novel view synthesis and 3D reconstruction has recently seen significant advancements, driven by several key models and approaches.

2. Related Work

The following methods represent the state-of-the-art in creating explicit or implicit 3D representations from a set of images or video frames:

- **Neural Radiance Fields (NeRF)**: NeRF [4] is a foundational method that represents a scene as a continuous volumetric function, mapping 3D coordinates $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ and viewing directions (θ, ϕ) to a volume density and view-dependent color. While NeRF excels at novel view synthesis, extracting a clean, manifold 3D mesh from the implicit representation often requires additional techniques, such as applying a Marching Cubes algorithm on a rendered density field.
- **3D Gaussian Splatting (3DGS)**: A more recent and computationally efficient technique, 3DGS [2] uses a collection of explicit, anisotropic 3D Gaussians (defined by position, covariance, and color) to represent the scene. It achieves state-of-the-art rendering quality and speed. However, like NeRF, the primary output is a rendering primitive, and a separate process, often involving an implicit surface reconstruction technique, is necessary to derive a watertight mesh.
- **COLMAP**: This is a traditional Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline. COLMAP first estimates camera poses and a sparse point cloud (SfM) and then generates a dense reconstruction (MVS). It provides a good initial geometric understanding and is often used as a preprocessing step to generate the required camera poses for learning-based methods like NeRF and 3DGS.
- **Neuralangelo**: This model combines the power of NeRF-like neural rendering with a zero-level set Signed Distance Function (SDF) to produce high-quality, geometric details. By implicitly representing the surface using an SDF, Neuralangelo [3] is designed explicitly for extracting high-fidelity 3D meshes that capture fine details from complex, real-world scenes.
- **SuGaR (Surface-Aligned Gaussian Splatting for Real-time 3D Reconstruction)**: SuGaR [1] represents an effort to bridge the speed and quality of 3DGS with the need for an explicit surface. It aligns the 3D Gaussians after regularization to an underlying mesh-like structure, often derived through Poisson Surface Reconstruction from a point cloud, aiming for a fast-reconstructed surface that is texture-rich.

3. Method

Our goal is to transform real-world video footage of a large structure into a block-based structure within a **Minecraft** world. The overall pipeline is divided into two primary stages: 3D reconstruction from video and subsequent mesh voxelization and conversion.

3.1. Video to 3D Mesh Reconstruction

The first critical step involves reconstructing a high-fidelity 3D mesh from the input video sequence.

We initially processed our video footage using **COLMAP**, with the intention of utilizing the estimated camera poses and dense point clouds, which are the standard input requirements for state-of-the-art neural rendering and reconstruction techniques.

We intended to utilize the following cutting-edge, out-of-the-box approaches, which promise superior geometric detail and mesh extraction, leveraging the extracted COLMAP parameters:

- **Neuralangelo**: Known for generating highly detailed 3D surfaces by optimizing a Signed Distance Function (SDF) implicitly represented by a Neural Radiance Field (NeRF) structure.
- **SuGaR (Surface-Aligned Gaussian Splatting for Real-time 3D Reconstruction)**: An approach that leverages the rendering speed of 3D Gaussian Splatting (3DGS) while simultaneously deriving an explicit, watertight surface.

We successfully executed the SuGaR pipeline using COLMAP-estimated camera poses derived from Google Earth imagery. The method converged and produced an explicit surface representation with recognizable architectural structure. However, the reconstruction exhibited fragmentation and noise in regions with inconsistent viewpoints, particularly around terrain and occluded areas. While the qualitative results were promising, SuGaR required substantially greater computation time and memory compared to alternative pipelines, making it impractical for large-scale experimentation within our project constraints. Consequently, we did not include SuGaR in quantitative evaluation and instead treat it as a qualitative reference point.

Alternative SfM Comparison We considered the use of **VGGT (Visual Geometry and Graphics Toolkit)** [5] as an alternative SfM pipeline, primarily due to its potential for faster execution compared to COLMAP. However, our qualitative assessment of preliminary results indicated that VGGT often produced a more sparse and less reliable set of camera poses and point cloud information. Given the strong connectivity and established compatibility of COLMAP's output with nearly all current neural reconstruction models, and the need for high geometric fidelity, we prioritized the use of COLMAP despite its longer processing time.

3.2. 3D Mesh to Minecraft World Conversion

The second stage converts the generated 3D mesh into a format compatible with **Minecraft**'s block-based architecture. This process is essentially a form of **voxelization**, converting the continuous surface geometry into discrete volumetric blocks.

To load the final 3D mesh (typically in the '.obj' format) into the game, we investigated two primary tools:

- **BlockPrintr Mod**: We initially attempted to use the `BlockPrintr` mod, which is designed to directly import meshes into the Minecraft game environment. However, we encountered significant technical difficulties. Despite manual adjustments to the mesh's orientation and scale, the imported structure consistently appeared with an incorrect rotation and orientation.
- **ObjToSchematic Utility**: We subsequently switched to the online tool `ObjToSchematic`. This tool converts meshes into the block-level file formats used by common Minecraft world editing tools: the `.schem` format (used by WorldEdit) and the `.litematica` format (used by Litematica).

The final output is a `.schem` file that accurately represents the reconstructed geometry using a customizable palette of Minecraft blocks, ready for in-game deployment.

3.3. Approach Evaluation

Due to time constraints, we were unable to define a numerical measure of video to Minecraft reconstruction. Due to the highly visual nature of our project, we evaluate our results purely on qualitative factors.

4. Experiments

4.1. Dataset and Preprocessing

Our experiments were conducted using imagery captured from Google Earth, focusing on prominent outdoor structures on the Emory University campus. Image sequences were extracted to provide sufficient viewpoint coverage for structure-from-motion and downstream reconstruction. Due to the synthetic nature of Google Earth imagery, camera viewpoints exhibit non-physical motion and view-dependent rendering artifacts, which introduce additional challenges compared to real-world photography.

Camera poses and sparse point clouds were estimated using COLMAP. These poses served as input to all subsequent reconstruction pipelines. We found COLMAP to produce the most stable and consistent camera estimates among the tools we evaluated, and it remains widely compatible with modern neural reconstruction methods.

4.2. Models and Results

We evaluated multiple structure-from-motion pipelines to assess their suitability for large-scale outdoor reconstruction using Google Earth imagery. Figure 2 shows a reconstruction generated using VGGT, which offers faster execution but produced unstable camera trajectories and distorted global geometry. In contrast, COLMAP produced more consistent camera poses and structurally coherent reconstructions.

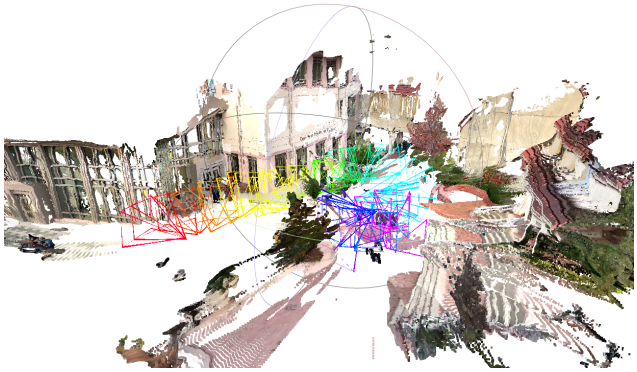

Figure 2. VGGT reconstruction of Candler School of Theology

Figure 3. Dense COLMAP reconstruction of Candler School of Theology

Figures 3 and 4 illustrate dense and sparse COLMAP reconstructions, respectively. The dense reconstruction captures major architectural features with reasonable fidelity, while the sparse point cloud highlights camera coverage and scene observability. Based on these qualitative comparisons, we selected COLMAP as the primary preprocessing step for all downstream reconstruction experiments, including neural and mesh-based pipelines.

- **COLMAP-Based Mesh Reconstruction**: As a baseline, we relied on COLMAP's multi-view stereo reconstruction to generate dense point clouds and meshes. While this approach provided reasonable global structure, the resulting meshes often lacked fine geometric detail and exhibited noise in regions with limited viewpoint coverage. Nevertheless, this pipeline proved computationally efficient and provided a reliable foundation for downstream voxelization.
- **Surface-Aligned Gaussian Splatting (SuGaR)**: We additionally experimented with SuGaR (Surface-Aligned Gaussian Splatting), a recent method that augments 3D

Figure 4. Sparse COLMAP reconstruction of Glenn Memorial Church.

Gaussian Splatting with surface regularization to enable explicit surface extraction. Using COLMAP-estimated camera poses from our Google Earth dataset, we successfully executed the SuGaR pipeline end-to-end on an NVIDIA A100 GPU. SuGaR converged and produced an explicit surface representation that captured recognizable architectural elements, including building facades and structural boundaries. Qualitatively, the reconstructed geometry was visually coherent in well-observed regions but exhibited fragmentation and noise in areas affected by inconsistent viewpoints, occlusions, and terrain variation. These artifacts are likely exacerbated by the synthetic, view-dependent rendering inherent to Google Earth imagery. While SuGaR produced promising qualitative results, it required substantially intensive computation time and memory. Training required multiple hours on high-memory GPUs, which limited our ability to perform extensive hyperparameter tuning or large-scale experimentation. As a result, SuGaR was not included in quantitative comparison and is treated as a qualitative reference point demonstrating the feasibility and limitations of surface-aligned Gaussian methods under our data constraints.

- **Mesh Voxelization and Minecraft Integration**: Meshes produced by the reconstruction pipelines were converted into Minecraft-compatible formats via voxelization. We evaluated two approaches: the BlockPrintr mod and the ObjToSchematic utility. Due to persistent issues with orientation and scaling when importing meshes directly into Minecraft using BlockPrintr, we adopted ObjToSchematic as our primary conversion tool. The voxelized outputs were exported as .schem files and loaded into Minecraft using standard world-editing tools. Visual inspection confirmed that the reconstructed geometry preserved the overall spatial structure of the original scenes, although fine geometric details were limited by voxel resolution and reconstruction quality.
- **Evaluation Protocol**: Due to the absence of ground-truth 3D geometry for large-scale outdoor environments and the visual nature of our final application, we evaluate all results qualitatively. We assess reconstruction quality based on:
  - Structural coherence of buildings and terrain
  - Presence of major architectural features
  - Visual noise and fragmentation
  - Suitability for voxelization and in-game representation

  Quantitative metrics such as PSNR, SSIM, or Chamfer Distance were not computed, as our primary objective was visual fidelity within a block-based environment rather than photometric or geometric optimality.

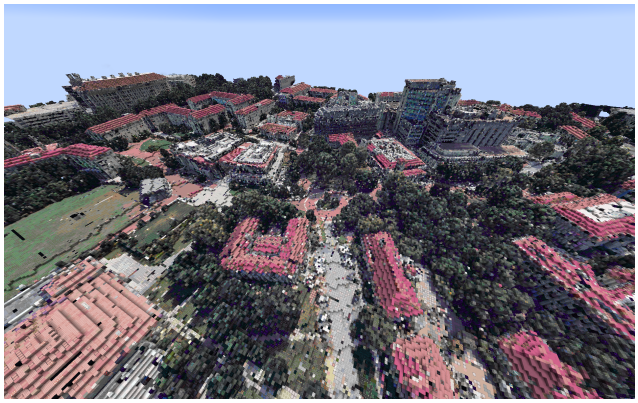

Figure 5. 3D reconstruction of Emory University from Google Maps data.

Due to the lackluster results from all tested models, we decided to approach our initial problem with the simplest solution: given the sheer amount of resources and hence the sheer amount of data availability at Google, we used a preexisting 3D reconstruction of Emory University from Google Maps/Earth. With some data extraction tools, we were able to load the model into Blender and perform some manual clean-up to import the final model to Minecraft, as seen in Figure 5.

5. Discussions

Our experiments highlight both the premise and limitations of modern video-to-3D reconstruction pipelines when applied to large-scale, synthetic outdoor imagery and downstream voxel-based applications. While recent neural reconstruction methods demonstrate impressive visual fidelity under controlled conditions, their performance and practicality vary significantly depending on data characteristics and computational constraints.

A central challenge in our work arises from the use of Google Earth imagery. Although visually realistic, Google Earth data violates several assumptions underlying state-of-the-art reconstruction methods, including consistent camera motion, physically accurate perspective, and photometric consistency across views. These violations manifest as fragmented geometry, noisy surfaces, and instability in poorly observed regions, particularly for learning-based methods that rely on dense and coherent multi-view supervision.

The SuGaR pipeline illustrates this tradeoff clearly. While SuGaR successfully produced explicit surface representations with recognizable architectural structure, it required substantial computational resources and exhibited sensitivity to view inconsistency and occlusion. In contrast, COLMAP-based reconstruction, although less detailed, proved more robust and computationally efficient for our dataset. This suggests that for large-scale, visually complex environments intended for coarse geometric representations, traditional structure-from-motion pipelines may offer a more favorable balance between reconstruction quality and practical usability.

Importantly, our end goal differs from many prior works in neural rendering and reconstruction. Rather than optimizing for photorealistic novel view synthesis or precise geometric accuracy, our objective is to generate a structurally faithful representation suitable for voxelization and deployment in a block-based environment. In this context, minor geometric inaccuracies and surface noise are often less critical than global structural coherence and computational scalability. Methods that prioritize speed, stability, and predictable outputs may therefore be more appropriate than highly expressive but resource-intensive neural models.

These observations suggest several directions for future work. First, incorporating further processing before and after COLMAP to get rid of unwanted point clouds via image segmentation/background removal or radius/bounding-box threshold. Second, incorporating additional regularization or geometry-aware constraints tailored to synthetic imagery may improve the stability of surface-aligned Gaussian methods. Third, hybrid pipelines that combine fast SfM-based geometry with lightweight neural refinement could offer a more practical middle ground. Fourth, exploring mesh/schema import approaches in Minecraft that would not cost prohibitive amounts of memory or building a pipeline to import several schemas at once and smoothly join them. Finally, exploring task-specific evaluation metrics aligned with voxel-based representations may better capture reconstruction quality for applications such as virtual environments and games.

Overall, our findings emphasize that the suitability of a reconstruction pipeline depends not only on its theoretical performance but also on its alignment with data characteristics, computational constraints, and downstream objectives.

6. Conclusions

In this paper, we explored multiple video-to-3D reconstruction pipelines with the goal of representing real-world architecture within the block-based environment of Minecraft. Using imagery from Emory University campus, we investigated the feasibility of converting large-scale outdoor scenes into voxelized structures suitable for in-game deployment.

Our experiments demonstrate that while modern neural reconstruction methods such as Surface-Aligned Gaussian Splatting can produce visually compelling surface representations, they come with significant computational costs and sensitivity to data inconsistencies, particularly when applied to synthetic imagery such as Google Earth. In contrast, traditional structure-from-motion pipelines based on COLMAP, although less detailed, provided more stable and computationally efficient reconstructions that were better aligned with our downstream voxelization requirements.

These results highlight an important practical insight. For applications that prioritize structural fidelity and scalability over photorealism, simpler and more robust reconstruction methods may be preferable to more expressive but resource-intensive neural approaches. Additionally, the challenges encountered with synthetic, view-inconsistent data underscore the need for reconstruction pipelines that are resilient to imperfect camera models and rendering artifacts.

Overall, our work demonstrates the feasibility of transforming real-world environments into interactive, block-based virtual representations and provides a comparative perspective on the tradeoffs between reconstruction quality, computational cost, and application-specific constraints. We hope this project serves as a foundation for future exploration into hybrid reconstruction pipelines and task-aware evaluation metrics for virtual environment generation.

References

[1] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering, 2023. 2

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 2

[3] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction, 2023. 2

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 2

[5] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025. 2