Heritage3DMtl: A Multi-modal UAV Dataset of Heritage Buildings for Digital Preservation

Rucha Shende

A thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Computer Science (Computer Science) at
Concordia University
Montréal, Québec, Canada

April 2024

© Rucha Shende, 2024

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: Rucha Shende

Entitled: **Heritage3DMtl: A Multi-modal UAV Dataset of Heritage Buildings for Digital Preservation**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Dr. Thomas Fevens	Chair
______________________________ Dr. Thomas Fevens	Examiner
______________________________ Dr. Dhrubajyoti Goswami	Examiner
______________________________ Dr. Tiberiu Popa	Co-supervisor
______________________________ Dr. Sudhir P. Mudur	Supervisor

Approved by ______________________________
Dr. Hovhannes Harutyunyan, Graduate Program Director
Department of Computer Science and Software Engineering

__________ 20 _____ ______________________________
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Heritage3DMtl: A Multi-modal UAV Dataset of Heritage Buildings for Digital Preservation

Rucha Shende

Unmanned Aerial Vehicle (UAV) technology has emerged as a transformative tool for 3D reconstruction, offering diverse applications in urban planning, heritage studies, infrastructure monitoring, and emergency response. Despite considerable progress, the field of heritage studies faces challenges due to the scarcity of real-world data tailored for heritage preservation. To address this gap, this thesis presents *Heritage3DMtl*, an extensive multi-modal dataset comprising 17 heritage buildings in Montreal, acquired using a UAV. The dataset includes images, estimated camera poses, and reconstructed 3D data (point clouds and meshes), providing great detail and diversity.

A Standard Operating Procedure (SOP) for data collection is provided, demonstrating the efficient use of low-cost consumer-grade UAVs to capture heritage buildings. This SOP serves as a replicable blueprint for future similar efforts. Various 3D reconstruction techniques are explored and experimented with using the dataset. Additionally, the dataset's applicability is showcased through the reconstruction of Level of Detail (LOD) models in alignment with the CityGML standard.

Furthermore, the integration of advanced reconstruction techniques, such as NeRF and Gaussian Splatting, has revolutionized the way we visualize and interact with building sites in digital environments. These techniques enable the generation of photorealistic renderings as well as interactive 3D models, enhancing our ability to study and interpret heritage buildings with unprecedented fidelity and detail.

In summary, this work contributes to the discourse on 3D heritage reconstruction by introducing an open-source dataset that enhances resources available to researchers and practitioners. *Heritage3DMtl* facilitates advancements in the field and serves as a valuable asset for digital preservation efforts, providing a comprehensive foundation for future research and innovation in heritage documentation and 3D reconstruction.

Acknowledgments

I would like to express my deepest gratitude to Dr. Sudhir Mudur, my supervisor, for giving me the opportunity to pursue this research. His unwavering support, guidance, and valuable feedback have significantly influenced the outcome of this work. I extend my sincere appreciation to Dr. Tiberiu Popa, my co-supervisor, for providing me with the necessary resources and for his valuable feedback and guidance throughout. My heartfelt thanks go to Dr. Kaustubha Mendhurwar for his exceptional mentorship. I am forever grateful for his unwavering support and encouragement throughout this journey.

I would like to mention that planning and executing drone flights to gather data for this work was an exhilarating and enjoyable learning experience. I feel fortunate to have contributed meaningfully, and I thank my supervisors for their support in facilitating this process.

I dedicate this thesis to my late father, whose dream was for me to pursue education abroad. Even though he is no longer with us, I know he is looking down on me with pride. I am profoundly grateful to my mother for her unwavering support, belief, love, and sacrifices. I am also thankful to my grandparents and my extended family.

I would also like to thank my family in Montreal for embracing me and providing me with a nurturing environment that made this journey feel like a home away from home. Finally, I am thankful to all my friends and labmates whose camaraderie and encouragement made my master's journey joyful; I am truly grateful for your presence in my life.

Contents

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**

1.1 Overview . . . 1

1.2 Contributions and Outline . . . 2

2 Background and Related Work **4**

2.1 3D Reconstruction Stages . . . 4

2.2 Geometric Representations . . . 6

2.2.1 Point Cloud . . . 6

2.2.2 Mesh . . . 6

2.3 Photogrammetry . . . 6

2.4 Neural Radiance Fields (NeRF) . . . 10

2.5 3D Gaussian Splatting . . . 10

2.6 CityGML LOD Models . . . 11

2.7 Polygonal Surface Reconstruction . . . 13

2.8 Relevant Computer Vision Techniques Used in our LOD Pipeline . . . 14

2.8.1 Use of Vision-Language Models (VLMs) . . . 14

2.8.2 Zero Shot Learning . . . 15

2.8.3 Zero Shot Image Segmentation . . . 15

2.8.4 Open-Set Object Detection . . . 15

2.9 Related Work on Heritage Dataset Creation . . . 16

2.9.1 LiDAR-Based Datasets . . . 16

2.9.2 Image-Based Datasets . . . 16

2.9.3 UAV-Based Datasets . . . 16

2.9.4 Synthetic Datasets . . . 17

3 Dataset Acquisition 19

3.1 Motivation . . . 20

3.2 UAV Data Collection . . . 21

3.2.1 Equipment and Setup . . . 21

3.2.2 Selection of Buildings . . . 21

3.2.3 Standard Operating Procedure (SOP) for Flight Planning and Control . . . 22

3.2.4 Quality Control . . . 22

3.2.5 Limitations . . . 24

3.3 Safety Brief . . . 25

3.4 Dataset Diversity and Characteristics . . . 25

3.5 Comparison of Proposed UAV Modality with Aerial LiDAR and Ground Data . . . 26

3.6 Conclusion . . . 28

4 Experiments: Dataset Processing and Applications 30

4.1 SfM + MVS Reconstruction . . . 30

4.2 Scene Representation . . . 32

4.2.1 Background and Related Work . . . 32

4.2.2 Implementation Details . . . 32

4.2.3 Results and Discussion . . . 33

4.3 Level of Detail 3 (LOD3) Modelling . . . 34

4.3.1 Background and Related Work . . . 34

4.3.2 Implementation Details . . . 35

4.3.3 Results and Discussion . . . 38

4.4 Conclusion . . . 39

5 Conclusions and Future Work 40

5.1 Introduction . . . 40

5.2 Contributions . . . 41

5.3 Future Work . . . 42

Appendix 43

References 44

List of Figures

1 Modalities/ Representations/ Applications 5

2 Point Cloud vs Mesh 7

3 3D Point Triangulation across multiple images (source: [1]) 8

4 Overview of NeRF training differentiable rendering approach. (source: [2]) 10

5 Conceptual difference between NeRF and Gaussian Splatting (source: [3]) 11

6 Level of Detail (LOD) 0-3 models represented in CityGML 3.0 (source: [4]) 12

7 Pipeline of PolyFit: (a) Input point cloud (b) Planar segments (c) Supporting planes of the initial planar segments (d) Supporting planes of the refined planar segments. (e) Candidate faces. (f) Reconstructed model (source: [5]) 13

8 Westmount City Hall: A sample from our dataset. 19

9 Screenshot depicting POI selection of Maison Shaughnessy using the FocusTrack feature of DJI Mini 3 Pro. 23

10 Sample images captured (left to right) in nadir, oblique and facade views. 24

11 Comparison of point clouds obtained from fusing LiDAR + Ground photos with those from UAV data. 27

12 Examples of camera locations estimated by Meshroom 31

13 Meshes generated from Meshroom (MVS), nerfacto and Surface-Aligned Guassian Splatting (SuGaR) 33

14 Novel views generated by 3D Gaussian Splatting and NeRF (nerfacto) 34

15 Result of Grounding DINO + SAM with text prompt: domes 35

16 Matching filtered key points across two views 36

17 Overview of the workflow for LOD3 generation 37

18 Sample results of generated LOD2 and LOD3 models on our dataset. 38

List of Tables

1 A comparative analysis of various studies that have undertaken data collection for the purpose of reconstructing heritage buildings. 17

2 Heritage3DMtl Dataset Statistics: Total UAV Images, Point Cloud Vertices, and Mesh Faces by Building 26

3 Year of establishment and a brief description of the buildings captured 29

4 Statistics of reconstructed LOD models 39

Chapter 1

Introduction

1.1 Overview

Recent advancements in computer vision, fueled by the rapid expansion of aerial digital imagery, have expanded the scope of 3D reconstruction research [6–8]. Unmanned Aerial Vehicles (UAVs) have played a pivotal role in this transformation, facilitating applications in building and infrastructure monitoring, urban planning, heritage preservation, and the integration of Geographic Information Systems (GIS) with virtual and augmented reality. UAVs are particularly valuable in scenarios where access to structures is limited and/or hazardous, such as post-natural disaster damage assessment [9, 10], bridge crack detection [11, 12], coastal environment mapping [13], survey of the coastal cliff faces [14], and autonomous drone delivery [15].

The adoption of photogrammetry and Neural Radiance Fields (NeRF) based scene 3D reconstruction has garnered significant attention in recent years, [16–20]. These techniques play a crucial role in preserving historic buildings, enabling visualization, reconstruction, as well as structural analysis [21–23]. By leveraging 2D images, these methods produce detailed and accurate 3D representations that capture the intricate details and semantic information of buildings.

Despite the immense progress in 3D reconstruction, there is a shortage of real-world data tailored for 3D heritage research. Existing datasets are either synthetic or are designed for large-scale city scene reconstruction. As a result, they do not contain the level of detail necessary for reconstructing individual buildings. Although aerial Light Detection and Ranging (LiDAR) data is increasingly available for urban scenarios, it is not suitable for structure-specific scenarios where images are intentionally captured to preserve geometric details. Our research addresses this gap. The fusion of integrating ground-based camera data with aerial LiDAR data is an approach that first comes to mind. This fusion is, however, challenging owing to the differences in perspective as well as scale. Although terrestrial laser scanning (TLS) provides high-resolution ground data, its deployment is constrained by high costs and logistical challenges. In this work, we advocate for UAVs as an accessible and cost-effective solution for 3D reconstruction of heritage buildings without compromising on data quality and detail.

In this work, we introduce *Heritage3DMtl*, an extensive dataset of heritage buildings in Montreal, Canada, acquired using a drone. This dataset is meticulously collected, capturing architectural features, intricate designs, and structural elements to provide comprehensive representations of the buildings' exteriors. It includes UAV images, camera poses, point clouds, and meshes, making it a valuable resource for research in vision, graphics, and structural analysis.

To facilitate the storage and exchange of virtual 3D city models, the City Geography Markup Language (CityGML) [24] has emerged as an open standardized international data model and exchange format. It defines four different levels of detail (LOD) 0-3 models. LOD3 model, in particular, captures detailed architectural elements (such as windows, doors, etc.) as well as semantics, making it suitable for cultural heritage reconstruction, documentation, architecture, and urban planning. We propose a novel component in existing workflows for automatic generation of LOD3 models of these heritage structures. Through this work, our goal is to encourage additional research in the realms of 3D heritage reconstruction and preservation.

1.2 Contributions and Outline

To summarize, our main contributions include:

1. *Heritage3DMtl Dataset*: A comprehensive and open-source dataset containing exterior appearance and geometric data of 17 heritage buildings in Montreal, Canada. This dataset, captured using a low-cost UAV, represents numerous architectural styles and offers detailed representations. It contains acquired UAV images, camera poses, point clouds as well as meshes of each building. To our knowledge, this dataset stands as the first such extensive data set to date in

this domain and aims to foster further research in digital heritage preservation and reconstruction.

2. *Low-Cost Capture Process*: We present an effective capture process for detailed structures, emphasizing the advantages of UAV-based data collection over ground cameras and/or LiDAR data. We discuss the design evolution of our acquisition method and the subsequent data processing pipeline.

3. *Utility Demonstrations*: We showcase the utility of both image and geometric representations in the dataset by generating novel views using NeRF and Gaussian Splatting techniques. Additionally, we propose a workflow for generating LOD3 models using open vocabulary zero-shot detection and segmentation.

The thesis is structured as follows: Chapter 2 provides background information and a review of the related work. Chapter 3 outlines the Standard Operating Procedure (SOP) we developed for data acquisition and provides details of the captured buildings. Chapter 4 presents experimental results and discusses the dataset's applicability in reconstructing lightweight polygonal LOD models. Finally, Chapter 5 presents our conclusions and outlines avenues for future research.

Chapter 2

Background and Related Work

In this chapter, we provide a brief overview of the background material related to this manuscript and review work related to 3D reconstruction and representation of the buildings, with emphasis on geometric representations. Additionally, we provide a comprehensive overview of existing datasets, focusing on imagery of buildings in heritage or urban scenarios captured for reconstruction purposes.

2.1 3D Reconstruction Stages

The process of converting 2D or 3D data of individual buildings or entire cities into digital formats suitable for different applications involves three primary stages:

1. **Capture Modality and Planning**: This stage entails choosing a collection of cameras and sensors to capture and digitize the physical environment at specific points in space. The sensors may comprise aerial or ground cameras for photogrammetry, LiDAR for precise distance measurements, or other specialized equipment designed to capture the shape, colour, and texture of environments. The choice of sensor affects the quality, type, and usability of the captured data. Planning involves deciding on sensor placements, the extent of coverage, and the resolution needed to adequately represent the target environment.

2. **A Representation Model**: This stage involves building a coherent and consistent model from sensed data. There exist various forms of representations, including point clouds, triangle meshes, parametric models from computer-aided design (CAD), or newer methods like NeRF [25] or 3D Gaussian Splatting

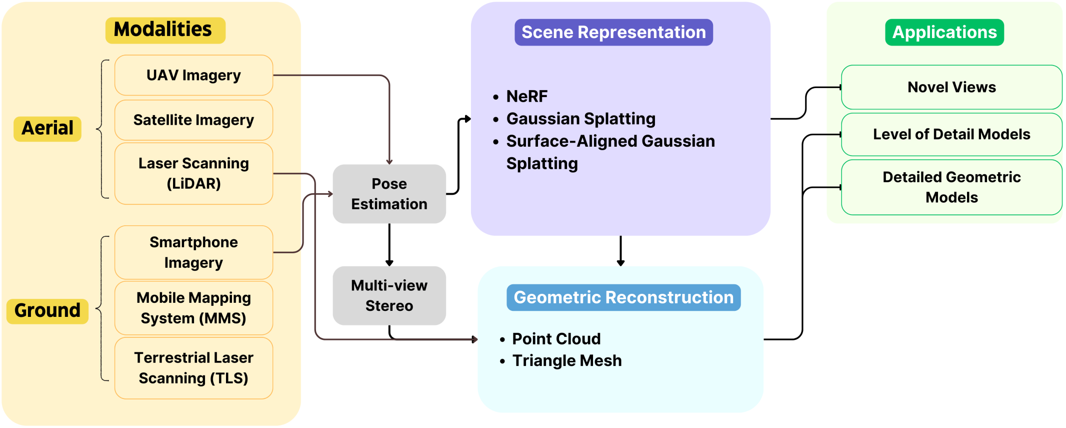


Figure 1: Modalities/ Representations/ Applications

[26]. The chosen representation significantly influences the fidelity, realism, and usability of the digital model. For instance, point clouds and triangle meshes provide geometric representations of the surface of objects, suitable for applications requiring accurate physical dimensions. On the other hand, NeRFs and 3D Gaussian Splatting offer volumetric scene reconstruction, which is better for capturing complex light interactions and textures in a scene. The intended purpose impacts the selection of a representation model, whether that's for visualization, simulation, or analysis.

3. **An Application Layer**: This stage focuses on utilizing the data across a broad spectrum of applications, including engineering for structural analysis, digital twinning, digital preservation, rendering, urban planning, architectural design, cultural heritage preservation, disaster management, real estate visualization, and the development of autonomous vehicles. These digital models are crucial in boosting efficiency, aiding sustainability initiatives, and offering immersive educational and entertainment experiences.

In today's rapidly evolving technological landscape, the array of available options for data capture, representation, and application is vast, as illustrated in Fig. 1. Given this complexity, it is crucial for modern heritage and building datasets to possess a high degree of versatility, capable of accommodating various representations and serving a wide range of applications. In Chapter 4, we provide comprehensive insights into our methodology for processing the acquired data, employing multiple representations to ensure thorough coverage. Additionally, we present the outcomes of experiments conducted to showcase the diverse applications of our dataset.

2.2 Geometric Representations

Point clouds and dense triangle meshes are two important structures for representing 3D geometry of objects and scenes, and techniques for deriving such structures from images have reached significant maturity.

2.2.1 Point Cloud

A point cloud represents a collection of data points in 3D space, where each point is defined by its cartesian coordinates (X, Y, Z). These points collectively depict the external surfaces of objects or scenes in three dimensions. While point clouds offer direct visualization and inspection of object shapes in 3D, they are often converted into more practical formats, such as polygon mesh models or CAD models through a process known as surface reconstruction. Point clouds play a pivotal role in building modelling by utilizing technologies like LiDAR, photogrammetry, and Building Information Modeling (BIM) [27–29].

2.2.2 Mesh

A mesh is composed of vertices (points in space), edges (line segments connecting pairs of vertices), and faces (polygons enclosed by edges). Meshes describe the surfaces of objects more completely by detailing their geometric and topological properties. They provide a more structured representation than point clouds, offering information on the connectivity between points. They accurately represent the physical dimensions of objects and environments, making them ideal for applications requiring high-fidelity models [30]. In the context of building reconstruction, mesh models are utilized to create water-tight representations of buildings with precise shapes and scales [31, 32].

2.3 Photogrammetry

Photogrammetry is a technique used to derive 3D geometric data from a given set of unordered 2D images or videos. It essentially reverses the process of photography, where depth information is lost because a 3D scene is projected onto a 2D plane. Photogrammetry extracts 3D measurements and spatial data from images using various software tools, both open-source and commercial. Examples of open-source software include Meshroom [33], COLMAP [34], OpenMVG [35], 3DF Zephyr Free, among others. In this work, we extensively employed Meshroom to estimate camera

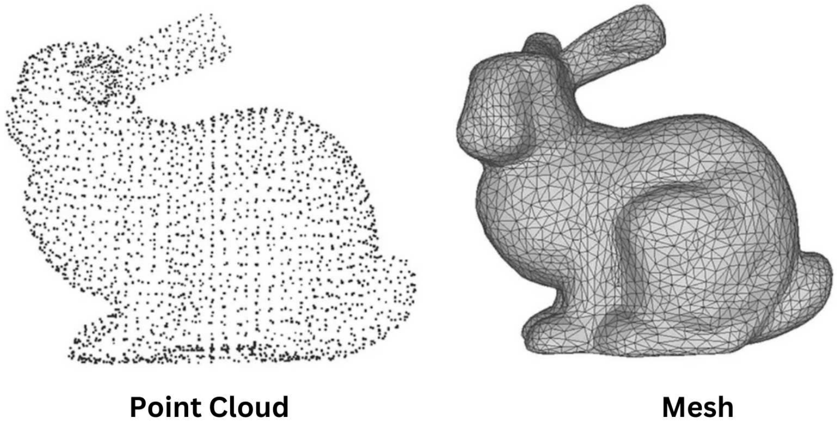


Figure 2: Point Cloud vs Mesh

poses and generate point clouds, dense meshes, and textured meshes. Meshroom offers a node-based graph editor and facilitates adjustment and visualization at each processing step. A typical photogrammetry pipeline for extracting point clouds and meshes works as below:

- **Feature Extraction:** This identifies and describes distinctive points or areas within each image, which can be accurately located across different images. These features are usually points of high contrast, edges, corners, or other significant visual markers within the image. Scale-invariant feature transform (SIFT) [36] is a common algorithm used, which extracts distinctive patches in an image with corresponding patches in another image irrespective of rotation, scale, or translation, and the output is a set of feature points for each image, each described by a unique descriptor (a vector of values). These features serve as reference points for comparing images to understand how they relate to each other in 3D space.

- **Image Matching:** Image matching refers to the process of identifying which images depict the same area or object from different viewpoints. This step organizes images into groups to enhance feature matching efficiency, particularly when reconstructing a large number of images. Identifying overlapping images reduces computational load during the next feature matching step.

- **Feature Matching:** Feature matching is the process of finding correspondences between the feature points extracted from different images. It involves comparing the feature descriptors to identify the same point in two different

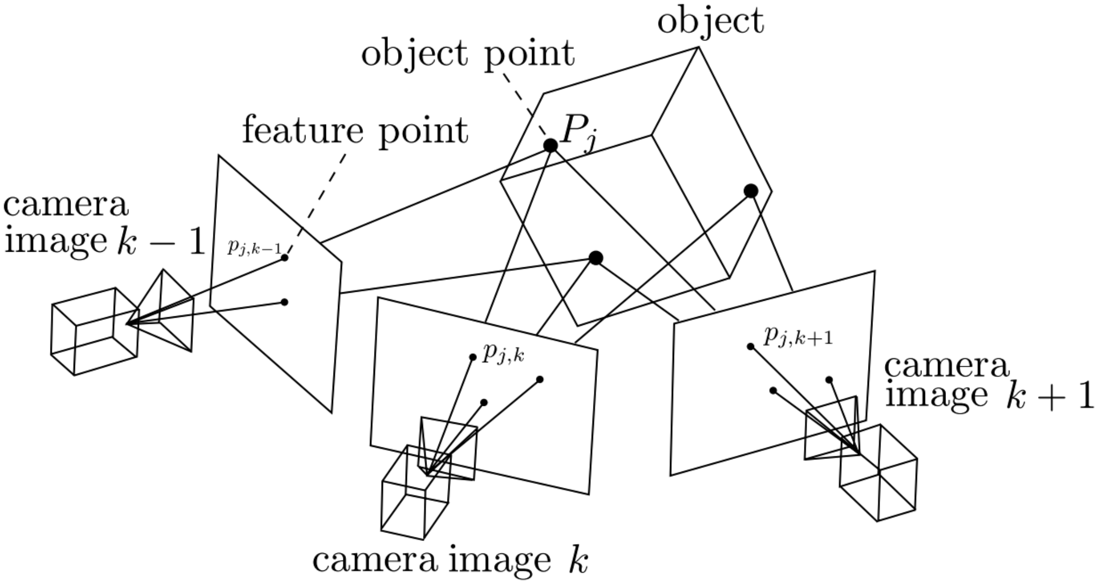


Figure 3: 3D Point Triangulation across multiple images (source: [1])

viewpoints. Algorithms, like nearest neighbour search, are used to find matches based on the similarity of descriptors. Robust methods, such as Random Sample Consensus (RANSAC), are then applied to eliminate outliers—incorrect matches likely due to repetitive patterns, changes in lighting, or occlusions. Accurate feature matches play a crucial role in precisely estimating the scene's geometry and the relative positions and orientations of the cameras, thereby enabling precise 3D reconstruction.

- **Structure from Motion (SfM):** It entails camera pose estimation and 3D point triangulation to create a sparse point cloud. Prior determines the position and orientation of each camera, while the latter calculates 3D coordinates of scene points observed across multiple images.

  Camera Pose Estimation: This process determines the extrinsic parameters - position and orientation (pose) of each camera when its corresponding image is taken. It determines the camera's placement and viewing direction in the 3D space for each image. Extrinsic parameters are mathematically represented by a transformation matrix combining a rotation matrix and a translation vector as shown in Eq. (1), which together describe the camera's placement and viewing direction in relation to a world coordinate system. Perspective-n-Point (PnP) algorithm [37], implemented within a RANSAC framework, is used to robustly estimate the camera pose, and also to handle outliers in the feature matching process.

$$\text{Extrinsics} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \tag{1}$$

3D Point Triangulation: Triangulation involves using the matched features (correspondences) between points across different images to 3D coordinates of these points in the scene. This is illustrated in Fig. 3. It uses camera projection matrices P and P' derived from camera poses estimated earlier. This method is based on epipolar geometry whose key components include epipolar lines, epipolar planes, and the epipole. An epipolar plane is defined by a 3D point and the centers of two cameras capturing the scene. The intersections of this plane with image planes produce epipolar lines, along which the corresponding image point must lie. The epipoles are intersection points of the line connecting the two camera centers (the baseline) with the image planes. To compute the 3D coordinate X of a point observed in two images, Eq. (2) is solved using singular value decomposition (SVD) [38].

$$\begin{bmatrix} P & x & 0 \\ P' & 0 & x' \end{bmatrix} \begin{bmatrix} X \\ \lambda \\ \lambda' \end{bmatrix} = 0 \tag{2}$$

where x, x' are 2D projections of X and λ, λ' are constants representing the point's position along the rays from the cameras. A sparse point cloud is obtained after all the 3D points in the scene are calculated. This serves as an initial, yet geometrically precise depiction of the scene's structure.

- **Multi-View Stereo (MVS):** MVS is a technique that uses the principles of triangulation to retrieve the depth value of each pixel for all cameras resolved by SfM. MVS algorithms generate a dense point cloud that represents the scene with greater detail compared to the sparse cloud from SfM. The dense point cloud and depth maps serve as a foundation for creating a 3D mesh. Lastly, textures are extracted from the original images and applied to the mesh. This involves unwrapping the mesh to create a 2D representation, onto which the photographic textures are projected and stitched together to maintain visual consistency and detail.

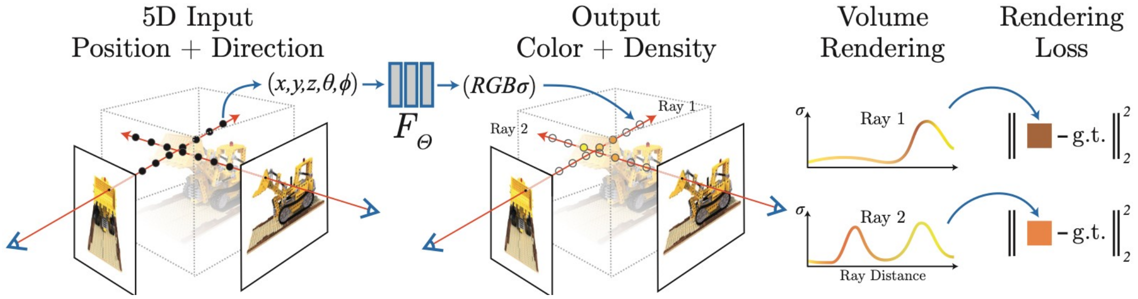


Figure 4: Overview of NeRF training differentiable rendering approach. (source: [2])

2.4 Neural Radiance Fields (NeRF)

NeRF, introduced in [2], represents a novel approach for synthesizing highly realistic images from new viewpoints based on a sparse set of input images of a scene. This technique optimizes a continuous, high-dimensional function that models the scene's radiance and density at any point in space. NeRF parameterizes this function using a multi-layer perceptron (MLP) network, which maps 5D coordinates (spatial location x, y, z and viewing direction (θ, ϕ) to colour (r, g, b) and density (σ). This allows for coherent renderings of complex scenes from viewpoints absent in the initial image set.

How do NeRFs differ from the classical 3D representation?
NeRF is a continuous, implicit representation that captures both the geometry (shape and structure) and appearance of the scene (colour and texture) within the network's weights. It, therefore, avoids the need for explicit geometric structures such as points or meshes. One key advantage of NeRFs is their ability to capture complex optical phenomena like occlusions, soft shadows, and scattering with photorealistic quality. These are aspects that classical representations, such as point clouds, often struggle to depict convincingly. However, it's worth noting that certain NeRF implementations do offer options for mesh recovery and export, bridging the gap between implicit and explicit representations.

2.5 3D Gaussian Splatting

Another innovative technique in 3D scene representation is 3D Gaussian Splatting [26]. Unlike NeRFs, which rely on volumetric rendering, Gaussian Splatting utilizes a novel approach based on 3D Gaussians or "splats." These splats, essentially 3D ellipsoids, can be rotated and stretched along any direction in space to capture the radiance and appearance of a scene efficiently (see Fig. 5). This method offers an

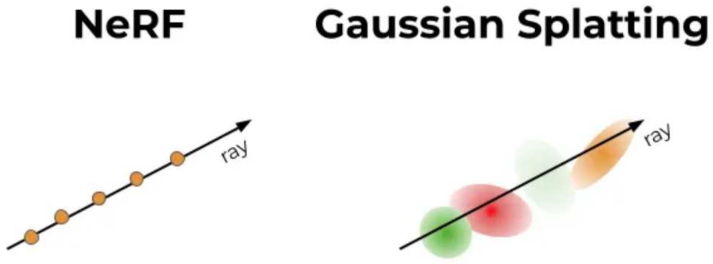


Figure 5: Conceptual difference between NeRF and Gaussian Splatting (source: [3])

alternative to NeRF, diverging significantly, offering a more efficient way towards real-time rendering at high quality, while maintaining competitive optimization times. It benefits from both explicit point-based representations and differentiable volumetric rendering. It begins by initializing 3D Gaussians from sparse point clouds (without normals) obtained via SfM. The Gaussians are described by their position (μ), a covariance matrix (Σ), opacity (α), and color (r, g, b) through spherical harmonics (SH). The formation of the radiance field representation is achieved through successive optimization steps (done using Stochastic Gradient Descent (SGD)) of the Gaussian parameters 3D Guassians were opted by the authors owing to their differentiability and ease of projection to 2D for rapid rendering. The optimization takes advantage of standard GPU-accelerated frameworks and custom CUDA kernels for efficient rasterization. By leveraging differentiable 3D Gaussian splatting and adaptive density control, this method achieves real-time rendering speeds significantly higher than existing NeRF implementations, without compromising on the visual quality.

In Chapter 4, we use both NeRF and Gaussian Splatting to generate novel views of the captured buildings. We compare mesh reconstruction quality with photogrammetry generated meshes.

2.6 CityGML LOD Models

CityGML, an open data model and XML-based format, serves as a standardized framework for storing and exchanging virtual 3D city models. It facilitates integration of various geospatial information about urban environments, including buildings, terrain, vegetation, and roads, among others. Through the utilization of CityGML, diverse stakeholders and software applications can effortlessly integrate multiple data sources, facilitating interoperability across various systems. One of the key features of

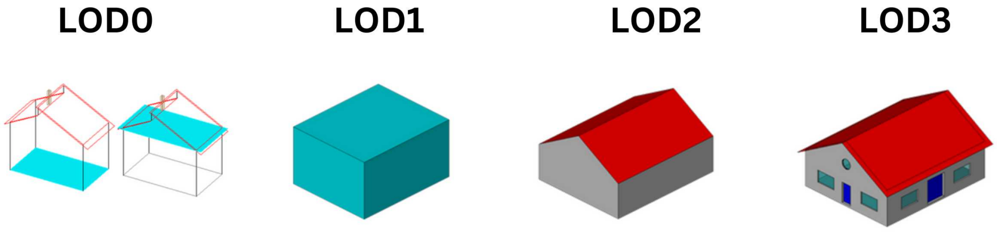


Figure 6: Level of Detail (LOD) 0-3 models represented in CityGML 3.0 (source: [4])

CityGML is the concept of LOD models. LOD models in CityGML refer to different representations of a 3D city model at varying levels of detail. These levels range from LOD0 to LOD3, each representing a different degree of complexity and accuracy in terms of geometry, semantics, and appearance. As we ascend the LOD hierarchy, the level of detail escalates.

- At the lowest level, LOD0 models consist of simple mass models with no geometric detail. It offers a fundamental depiction of the city's layout without specific details about buildings or other elements within the urban landscape.

- LOD1 model represents a rough block model without roof structures, suitable for general city models. This includes the representation of building footprints and the extrusion of these footprints to indicate building heights. This model provides a more realistic depiction of the city's structures, allowing for basic visualization and analysis.

- As we progress to LOD2, additional details are incorporated into the city model. In addition to building footprints and heights, LOD2 models include basic roof structures, such as simple gabled or hipped roofs. This level of detail enhances the visual realism of the city model and enables more advanced analysis, such as solar potential assessments.

- Advancing up the hierarchy, LOD3 models incorporate geometric, topological, and semantic elements. These models are vital for capturing intricate features of heritage structures, including the architectural details, windows, doors, and textures. Such details are crucial for preservation and documentation [39].

LOD models prove especially beneficial in depicting heritage buildings during risk and

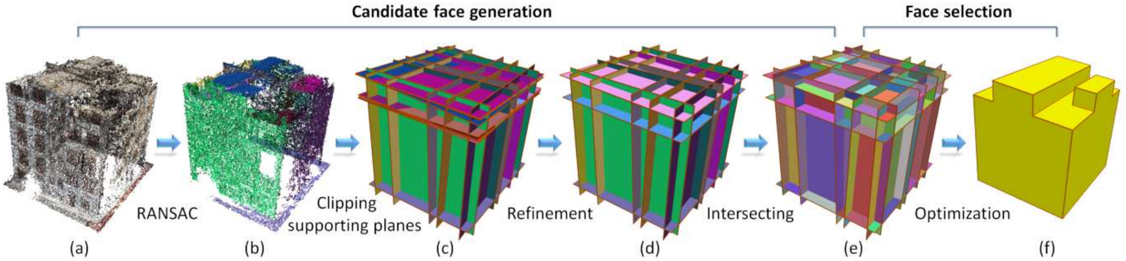


Figure 7: Pipeline of PolyFit: (a) Input point cloud (b) Planar segments (c) Supporting planes of the initial planar segments (d) Supporting planes of the refined planar segments. (e) Candidate faces. (f) Reconstructed model (source: [5])

damage assessments. They enable the incorporation of macro-elements and various feature types associated with damage mechanisms, thereby enriching the description and analysis of structural conditions [40]. Additionally, integrating Heritage Building Information Modeling (HBIM) with CityGML standards facilitates a multiscale 3D GIS approach for damaged cultural heritage. Furthermore, this integration enables a unified representation of building elements across various LODs [41]. In Chapter 4, we demonstrate a pipeline for generating LOD2 and LOD3 models from point clouds generated from our dataset.

2.7 Polygonal Surface Reconstruction

The reconstruction of 3D models from noisy or incomplete point clouds to achieve high-fidelity representations remains a significant challenge in computer vision and graphics. PolyFit: Polygonal Surface Reconstruction from Point Clouds [5] introduces a novel framework for reconstructing lightweight, polygonal surfaces from point clouds. It is based on hypothesizing and selection strategy and focuses on obtaining piecewise planar objects specifically for man-made objects such as buildings.

The framework begins by generating candidate faces from the point cloud. This process involves plane extraction using a RANSAC-based primitive detection method proposed in [42], which identifies initial planar segments. These segments undergo a refinement step aimed at merging similar planar segments to address issues caused by noise and outliers, thereby reducing computational complexity and avoiding thin, non-manifold, and degenerate faces in the final model.

Once refined, the supporting planes are pairwise intersected to generate a large set of candidate faces, encompassing potential parts of the final polygonal model. The subsequent step involves selecting the optimal subset of candidate faces to form the final model. This selection process is formulated as a binary linear programming problem, with the objective function comprising three energy terms: data-fitting, model complexity, and point coverage.

The data-fitting term ensures that selected faces closely match the input point cloud, while the model complexity term encourages simpler models with fewer sharp edges. Additionally, the point coverage term aims to minimize uncovered regions on the model. Hard constraints are enforced to ensure that the chosen subset of faces forms a manifold and watertight model.

Through this optimization process, the framework produces a polygonal surface that accurately represents the geometry of the object while remaining lightweight and free of gaps or unnecessary complexity. Given its effectiveness with building point clouds, we leverage this framework to generate LOD2 models, as described in Chapter 4.

2.8 Relevant Computer Vision Techniques Used in our LOD Pipeline

2.8.1 Use of Vision-Language Models (VLMs)

Vision-language models (VLMs) represent a crucial example of multimodal artificial intelligence (AI), facilitating the interpretation and generation of information by combining visual (images) and natural language (textual) data. This sophisticated approach involves the model's vision component recognizing visual elements in images, while the language segment processes textual information. Both the aspects, including object identification and the image's geometric configuration, are intricately mapped to each other. For instance, upon detecting an apple in a photograph, the model associates this visual element with corresponding terms found in textual descriptions. Typically, VLMs are initially trained on extensive multimodal datasets gathered from the web, comprising matching pairs of images/videos and text. They excel at tasks demanding an understanding of both images and text, such as object detection based on a text prompt, image captioning, visual question answering, and text-to-image generation. In our LOD3 construction pipeline, we employ pre-trained VLMs to identify and segment substructures such as windows and doors in buildings, utilizing zero-shot and open-set concept generalization methods, as further described below.

2.8.2 Zero Shot Learning

Traditional machine learning approaches heavily rely on extensive labelled datasets for model training, which becomes impractical or impossible when such labels are scarce or unavailable. Zero Shot Learning (ZSL) overcomes this limitation by enabling models to recognize objects or patterns they have never encountered during training. ZSL utilizes transfer learning to glean knowledge from one domain (seen classes) to another (unseen classes). The essence of ZSL lies in its capability to leverage the semantic relationships between known and unknown categories, often using attributes or descriptions to bridge the gap. This approach not only mitigates the challenge of data scarcity but also enhances the model's generalization capabilities, proving instrumental in solving problems within the domains of image and pattern recognition.

2.8.3 Zero Shot Image Segmentation

Zero-Shot Image Segmentation denotes the capability of a model to segment images into meaningful parts or categories it has never seen during training. Developed by Meta AI, Segment Anything Model (SAM) [43] serves as a foundational model for image segmentation, facilitating zero-shot transfer to new image distributions and tasks by interpreting prompts specifying what to segment within an image. SAM achieves this by comprehending the context and attributes of objects through their visual features and associated textual descriptions, allowing it to segment images with objects it has never explicitly been trained on. Trained on a diverse dataset of over one billion masks, SAM generalizes well to new types of objects and images without requiring additional training.

2.8.4 Open-Set Object Detection

Open-set concept generalization refers to a model's ability to recognize and classify inputs belonging to classes not seen during training, categorizing them as either unknown or belonging to a novel category. Open-Set Object Detection extends this capability to detecting and localizing objects within images or video frames, while also recognizing when detected objects belong to classes absent from the training dataset. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection [44] is a novel open-set object detection model that integrates DINO [45], a Transformer-based detector, with grounded pre-training. This enables it to detect objects described by textual inputs in human language, including category names and referring expressions. Grounding DINO uses pre-trained VLMs to realize complex scenes and detect objects by grounding textual descriptions to visual elements. VLMs play a vital role in training Grounding DINO, providing semantic understanding key to link visual data with textual descriptions, thereby enabling open-set object detection.

In Chapter 4, we demonstrate how we utilize a combination of SAM and Grounding DINO to improve our workflow for reconstructing LOD3 models.

2.9 Related Work on Heritage Dataset Creation

In this section, we review available datasets and case studies aimed at reconstructing heritage or urban buildings. We organize the discussion into subsections based on the type of data acquisition method: LiDAR-based datasets, image-based datasets, UAV-based datasets, and synthetic datasets.

2.9.1 LiDAR-Based Datasets

LiDAR Airborne Laser Scanner (ALS) data has been extensively used for 3D building reconstruction for the past two decades [46–48]. Yet, it has limitations in capturing detailed facade information and complex roof structures. For instance, Building3D [49] is the largest building modelling urban-scale dataset, consisting of point clouds, meshes, and wireframe models generated from LiDAR data but does not provide the level of geometric detail required for accurate reconstruction of building exteriors.

2.9.2 Image-Based Datasets

Recent research trends show an increased use of aerial images or images captured from mobile devices for structure-aware 3D reconstruction of buildings. [50–53]. However, to our knowledge, there is a lack of publicly available large datasets specifically captured for the reconstruction of heritage buildings or even modern buildings.

2.9.3 UAV-Based Datasets

Numerous studies have explored heritage structures using UAVs and other methods. However, these studies often focus on data acquisition from only one or two sites. For instance, Kyriou et al. [54] showcase the synergistic use of UAV and TLS data for documenting cultural heritage sites affected by geo-hazards, but they cover only a

Table 1: A comparative analysis of various studies that have undertaken data collection for the purpose of reconstructing heritage buildings.

Reference	**Type**	**Diversity**
Klapa et al. [58]	UAV+TLS+GNSS	2 buildings
Xu et al. [55]	UAV+TLS	1 building
Kyriou et al. [54]	UAV+TLS	1 building
Jo et al. [56]	UAV+TLS	1 building
Andaru et al. [57]	UAV + TLS	2 buildings
Luhmann, et al. [59]	UAV + TLS	2 buildings
Mwangangi et al. [29]	UAV	2 small scenes + 1 building
Themistocleous, Kyriacos, et al. [60]	UAV	1 building
Samadzadegan, et al. [61]	UAV	1 building
Karachaliou, Eleni, et al. [62]	UAV	1 building
Murtiyoso, et al. [63]	UAV	2 buildings
Heritage3DMtl(Ours)	**UAV**	**17 buildings**

single site, the hanging Holy Monastery in Greece. Xu et al. [55] present just a single case study of the Liao Family Temple in China with only 45 images captured via UAV. A similar study by Jo et al. [56] applied TLS and UAV for digital documentation, again of only a single temple, the Magoksa Temple in Gongju, Republic of Korea. Mwangangi et al. [29] explore the potential of UAV photogrammetric point clouds for building facade detection and 3D reconstruction with data captured in three European countries with two areas covering multiple buildings, typically urban, but with respect to heritage buildings, only a city hall building was captured with overlapping oblique and nadir views. Andaru et al. [57] collected the data for two architectural heritage buildings in Indonesia. Tab. 1 summarizes different heritage-related case studies using multi-sensor data acquisition reported in the literature.

2.9.4 Synthetic Datasets

While various large-scale synthetic datasets like Syndrone [64] and Airloc [65] are available, they lack the fidelity to real-world data essential for heritage research. While one cannot ignore other huge datasets like the OMMO [19], MegaNeRF [53], MatrixCity [20], UrbanScene3D [66], UAVStereo [67], Mid-Air [68], STPLS3D [69] and SensatUrban [70], even if these contain a wide range of scenes and objects, they are really not suitable for the specific and detailed reconstruction of individual buildings with complex external structures.

While existing datasets offer valuable insights, there remains a need for large-scale, publicly available datasets specifically tailored for heritage building reconstruction. Such datasets would facilitate the development and evaluation of novel reconstruction algorithms and contribute to the preservation of cultural heritage.

Chapter 3

Dataset Acquisition

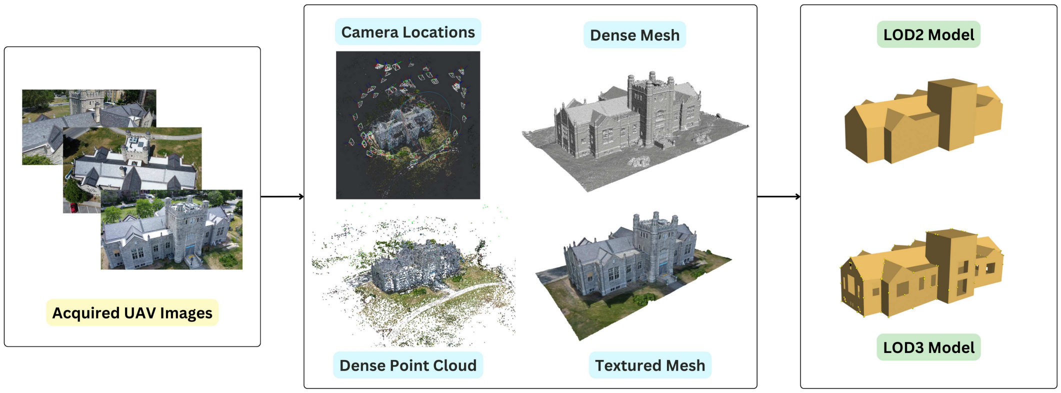


Figure 8: Westmount City Hall: A sample from our dataset.

In this chapter, we discuss the motivation behind creating *Heritage3DMtl*, our dataset and the choice of UAV as our data capture modality. Further, we delve deeper into the details of the image-capture process by laying out an SOP. This SOP allowed us to control the quality of our dataset. We also provide a safety brief and discuss the challenges faced during data collection. Further, we provide details of the buildings captured within our dataset. Fig. 8 illustrates the content of our *Heritage3DMtl*. Finally, we compare point clouds obtained on our data with publicly available LiDAR point cloud data to justify the choice of UAV as an ideal candidate for data acquisition.

3.1 Motivation

Montreal is renowned for its blend of classic and modern architecture. Yet, despite its wealth of architectural legacy, there remains a significant dearth of 3D datasets dedicated to heritage documentation and preservation.

Modern high-resolution and user-friendly smartphone cameras offer exceptional image capture capabilities in various conditions, making them a top choice for acquisition tasks. They excel particularly in capturing heritage artifacts, such as small statues when individuals have ample ground space to take photographs from optimal distances and angles [71]. However, many heritage buildings present a challenge owing to their height and limited available ground space, obstructed by structures, foliage, or water bodies.

While Terrestrial Laser Scanning (TLS) is a highly accurate method for obtaining detailed 3D data, it is hindered by two significant constraints. Firstly, the high cost associated with specialized equipment and expertise needed to operate it. Secondly, the logistical challenge related to the deployment of physical markers needed for data integration [72]. These markers demand considerable time and effort to set up, especially in large or complex sites. Furthermore, placing these markers can be impractical or restricted in environments with difficult terrain or sensitive historical locations, complicating the scanning process and adding to the complexity of aligning and integrating the scanned data [73, 74]. These factors make TLS less accessible for projects with limited budgets or those situated in challenging environments.

On the other hand, city-wide public data captured using LiDAR, satellite imagery, or aerial data from airplanes is available. While LiDAR yields point cloud data, it lacks colour and texture, necessitating the fusion of data from other aerial imagery. However, it introduces its own challenges, especially the absence of necessary facade details of heritage sites. Satellite or aerial images often lack detail, offer limited perspectives, and are prone to strong shadows due to their top-down view and distant range.

UAVs, notably mini-drones, are ideal devices for medium-sized buildings. They are low-cost, simple to operate, and loosely regulated in most jurisdictions. They can capture images from different altitudes, positions, lighting/weather conditions, and camera orientations. Integrating UAV technology in generating datasets of architectural heritage thus stands as an effective approach, combining detailed data capture with cost efficiency and operational flexibility.

Given the scarcity of available data, our work aims to bridge this gap by generating an extensive multi-modal dataset of Montreal's heritage buildings captured using a UAV. Such datasets, which include both 2D and 3D data, contribute greatly to the documentation, preservation, analysis, and modelling of historic structures.

3.2 UAV Data Collection

3.2.1 Equipment and Setup

We used two different consumer-grade UAVs, DJI Mini 2[1] and DJI Mini 3 Pro[2] for data capture. Both UAVs are categorized as mini-drones and are known for their compact size making them highly portable and easy to maneuver. They weigh less than 250 grams and do not need to be registered or need a pilot license to fly in Canada. The DJI Mini 2 is equipped with a 1/2.3 inch CMOS sensor capable of capturing 12MP effective pixels, while the DJI Mini 3 Pro features a 1/1.3-inch CMOS sensor capable of producing superior High Dynamic Range (HDR) images of up to 48MP. At the time of purchase, the DJI Mini 3 Pro stood out as the most advanced mini-drone model, featuring multi-direction obstacle sensing capabilities and a focus track functionality. We used the Mini 2 to capture three of the total seventeen buildings, while the rest were captured using the Mini 3 Pro.

3.2.2 Selection of Buildings

To ensure a representative dataset, we aimed to select heritage buildings with diverse architectural styles. We conducted a thorough evaluation of potential buildings, considering factors such as historical significance and architectural uniqueness. Additionally, we used Google Earth to assess occlusions surrounding the buildings. This step was crucial in identifying and excluding buildings where dense foliage or adjacent structures might obscure the views of the facades.

[1] https://www.dji.com/ca/mini-2

[2] https://www.dji.com/ca/mini-3-pro

3.2.3 Standard Operating Procedure (SOP) for Flight Planning and Control

Upon arriving at the location, we conduct a comprehensive on-site inspection to identify obstacles near each facade that may interfere with the image acquisition process. The home location with the fewest obstacles is then selected. Prior to takeoff, all necessary technical checks and calibrations are performed following the UAV user manual to guarantee a safe operation.

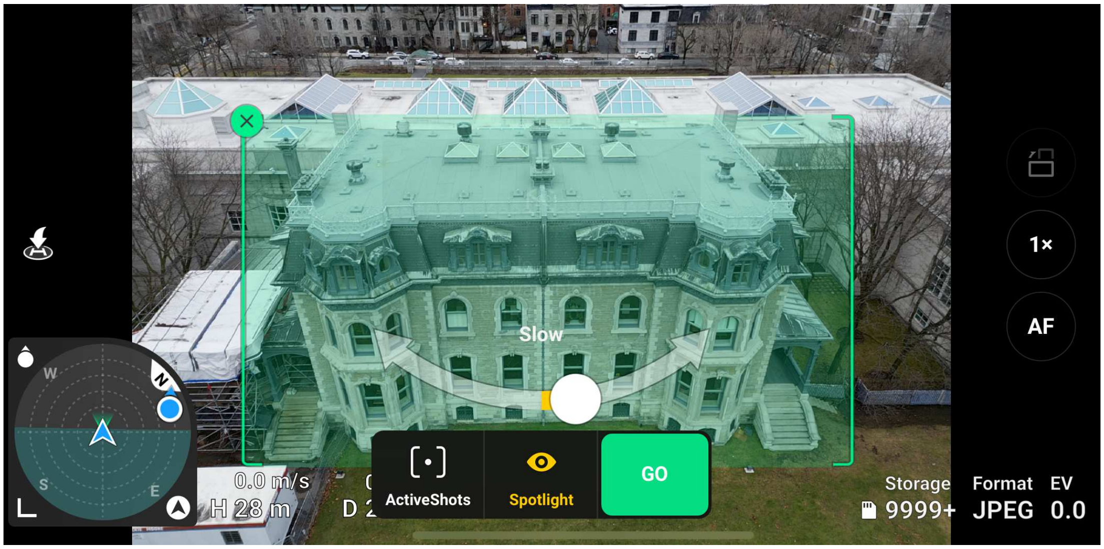
Capturing Nadir Images: We fly the UAV to the top of the building and set the gimbal's pitch angle to -90° while adjusting the yaw angle accordingly. We then rotate the UAV in 90° increments to capture images from four different perspectives at the same height. Due to height-restricted zones, we are not able to capture nadir images for some buildings. For such buildings, one possible solution is to fuse LiDAR point cloud data with the point cloud obtained on our UAV data.

Capturing Oblique Images: We fly the UAV above the highest obstacle, typically a tree, and position it to capture oblique images of the entire building. We adjust the gimbal angle between 0 and -90° to achieve an optimal perspective. The focus track functionality is activated, with the entire building selected as the point of interest (POI) as illustrated in Fig. 9. We then maneuver the UAV in a circular motion around the building at a suitable speed, capturing images with a 70-80% overlap between consecutive frames. This process is repeated multiple times with variations in height and angles to capture the building comprehensively.

Capturing Facade Images: We carefully descend the UAV to an appropriate distance from each facade, ensuring that the front exterior of the building is within the frame. We adjust the gimbal angle as necessary to capture images at different heights, maximizing the coverage of architectural details. This process is repeated for all accessible facades of the building. Although tree occlusions sometimes disrupted full facade captures, we focused on obtaining a higher number of oblique images by flying the UAV at varying altitudes to maximize the collection of oblique perspectives.

3.2.4 Quality Control

In order to ensure a more accurate and robust reconstruction of all the architectural facades, images must be captured with higher overlap and at different altitudes. [75–77]. The FocusTrack feature set is leveraged for intelligent tracking of the



Storage 9999+
Format JPEG
EV 0.0

Figure 9: Screenshot depicting POI selection of Maison Shaughnessy using the FocusTrack feature of DJI Mini 3 Pro.

building structures, facilitating the capture of a substantial number of images with increased overlap. Fig. 9 depicts how a POI is selected to track a building. The UAV automatically flies around the selected POI, allowing the flyer to capture shots at regular intervals at the same height. We employed the image capture process at varied distances and captured a wide array of nadir, oblique, and orthographic images, thereby augmenting the accuracy of the intrinsic and extrinsic camera parameters crucial for the reconstruction process. Fig. 10 illustrates a collection of images showcasing various perspectives of three buildings.

Nadir images, captured directly overhead, offer views of the building's roof, layout, and surface details. When stitched together with other images, they provide a consistent base for aligning and orienting other imagery, contributing to the accuracy and precision of reconstruction. Oblique images captured from various angles around the structure provide a more comprehensive view of the building, revealing intricate surface textures and elements that might not be fully captured in nadir shots. Facade view images facilitate the documentation of architectural features, such as windows, doors, and other structural elements. They facilitate generation of high-resolution texture maps for the surfaces of 3D models.

Data Preview and Assessment: We preview the collected data for image quality and comprehensiveness of coverage. Structures, regions, UAV locations and angles are identified as needed for additional captures. This is repeated until we are satisfied

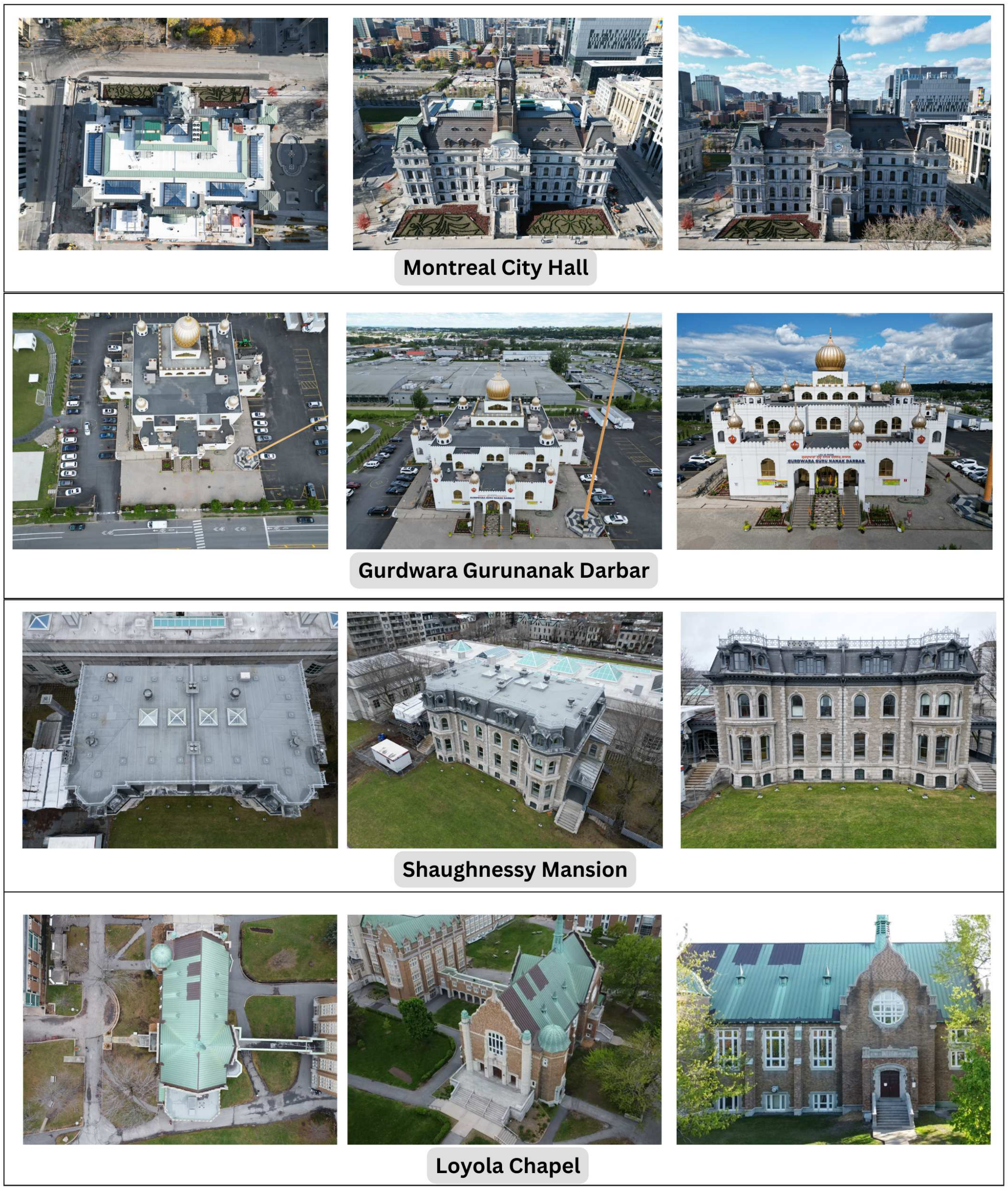


Figure 10: Sample images captured (left to right) in nadir, oblique and facade views.

that the UAV has covered all accessible regions of the building's exteriors. In practice, this step was only needed in 3 of the 17 captures that we did. All acquisition sessions took between 30 and 40 minutes.

3.2.5 Limitations

During the dataset collection process, we encountered various environmental and natural challenges like high winds, varying lighting conditions, shadows, temperature extremes, visitor crowds, regulatory restrictions, battery life, power lines and tree occlusions. To navigate through these obstacles, we strategically scheduled shoots during periods of calmer weather and low crowd conditions. Although tree occlusions disrupted facade image capture often, we focused on obtaining a higher number of oblique images by flying the UAV at varying altitudes to maximize the collection of oblique perspectives. Despite these challenges, the flexibility in our approach enabled us to preserve the dataset's integrity and comprehensiveness.

3.3 Safety Brief

Despite their advanced capabilities, UAVs can still pose a danger to human life and property, so we took great care to safeguard our acquisition sessions. We always conduct a thorough assessment of the weather conditions for the entire week; days with either sunny or cloudy weather are chosen. We also take into consideration the wind speed, preferring days with wind speeds up to 10 km/h and wind gusts not exceeding 25 km/h. These thresholds result from careful observation to ensure stable images despite the DJI Mini 3 Pro's wind resistance of up to 57 km/h. On sunny days, we plan the shoot around noon to avoid shadows and maintain consistent lighting conditions across all facades. Since heritage buildings are often also tourist attractions, we conduct visitor logistics to select dates and times when not many people are around.

We adhered to the regulations that govern the flight heights of UAVs. These regulations determine the altitudes at which UAVs can operate and impact the viewpoints from which buildings can be captured. Simultaneously, in strict observance of safety and legal guidelines, we ensured that our UAV operations were conducted at a safe distance from restricted zones, such as airports. Thus we guaranteed that our data collection process was legally compliant.

3.4 Dataset Diversity and Characteristics

Our meticulous data collection process involved systematic UAV flyovers, ensuring thorough coverage and capturing the intricate details and structural elements of each building from various angles. To the best of our knowledge, our dataset is the largest real-world UAV dataset of heritage buildings. It encompasses a total of 17 buildings from mid-19th to early 20th-century heritage structures in Montreal that represent a range of architectural styles, from Neo-Gothic churches and Victorian-era mansions to Beaux-Arts structures. Additionally, it includes intricately designed Hindu, Sikh, and Jewish temples and a Mosque, reflecting the city's historical, cultural, and religious evolution. These structures are detailed in Tab. 3. The dataset statistics with the total number of UAV images, number of vertices in reconstructed dense point clouds and number of faces in reconstructed meshes is given in Tab. 2. It stands as a valuable asset for research and applications in structural engineering, cultural heritage preservation, and city modelling, among others.

Table 2: Heritage3DMtl Dataset Statistics: Total UAV Images, Point Cloud Vertices, and Mesh Faces by Building

Building	**# Images**	**# Vertices**	**# Faces**
Montreal City Hall	247	791 K	1.60 M
Clock Tower	219	234 K	528 K
Saint Joseph's Oratory	243	549 K	1.09 M
Loyola Chapel	181	438 K	874 K
Holy Ghost Ukrainian Catholic Church	101	378 K	755 K
Westmount City Hall	65	1.01 M	2.02 M
St. George Antiochian Orthodox Church	66	740 K	1.48 M
Maisonneuve Market	394	914 K	1.34 M
Maison Louis-Hippolyte Lafontaine	236	972 K	1.94 M
Shaughnessy Mansion	150	687 K	1.37 M
St. Thomas More Church	156	827 K	1.39 M
Resurrection Chapel	301	657 K	1.39 M
Murugan Temple	202	741 K	1.63 M
Hindu Mandir	246	647 K	1.29 M
Gurdwara Guru Nanak Darbar, LaSalle	170	1.06 M	2.11 M
Congregation Shaar Hashomayim	117	800 K	1.67 M
Islamic Center of Quebec - El Markaz Islami	174	555 K	1.26 M

3.5 Comparison of Proposed UAV Modality with Aerial LiDAR and Ground Data

We compare the UAV modality against publicly available aerial LiDAR data and ground-level smartphone images. For this experiment, we select three buildings with distinct architectural styles: Saint Joseph's Oratory, Montreal City Hall and Gurdwara Gurunanak Darbar.

The LiDAR data for Montreal is available as LAZ tiles for individual neighbourhoods

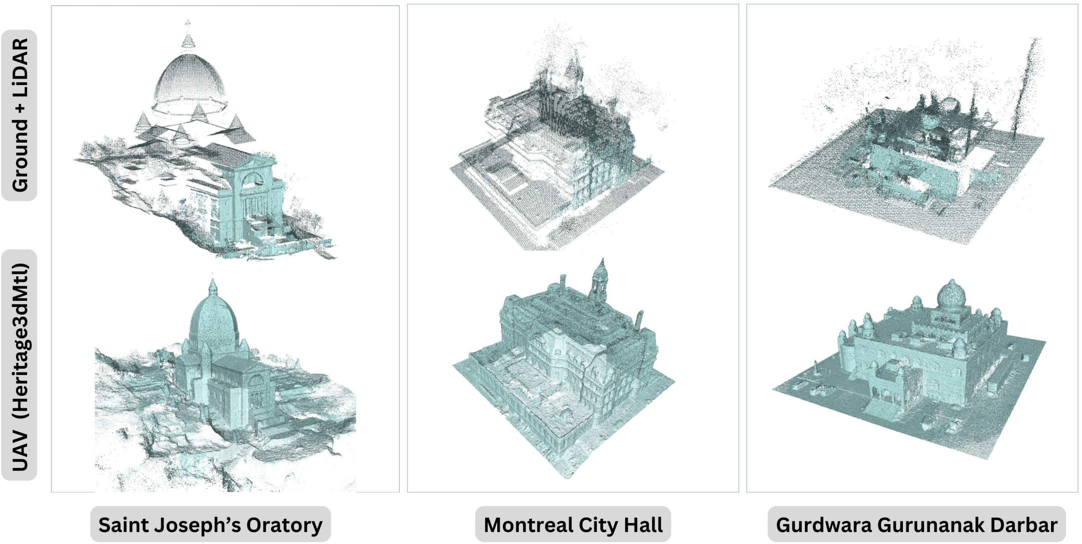


Figure 11: Comparison of point clouds obtained from fusing LiDAR + Ground photos with those from UAV data.

and boroughs. For data processing, we use CloudCompare, an open-source 3D point cloud processing software. We utilize its zooming, panning and rotation functionalities to navigate precisely to our target site. Then, we crop and refine the data with noise reduction to obtain a point cloud of a building without colour/texture information.

To obtain the ground-level images, we use an iPhone 13 Pro equipped with a 12 MP sensor, 1.9µm pixels, and a 26 mm equivalent f/1.5-aperture lens. Although the device also features a LiDAR scanner, it can only detect objects up to 5 meters. It is worth noting that capturing images from unreachable ground locations, such as behind the Saint Joseph's Oratory, is not possible. However, we capture an equivalent number of images to those obtained by UAV, with an overlap of $\approx 70 - 80\%$. We generate point clouds using Meshroom. Despite the significant overlap, SfM has difficulty properly aligning multiple cameras and estimating accurate camera orientation for several images, resulting in incomplete and inaccurate reconstructions.

Finally, we align and merge these point clouds with LiDAR point clouds in CloudCompare. The resulting point clouds are qualitatively compared, and it is evident that even after the fusion, significant information is missed in the facade regions. In comparison, by carefully choosing the flight paths for the UAV, the resulting point clouds are more complete and visually better. A qualitative comparison is shown in Fig. 11.

3.6 Conclusion

In this chapter, we embarked on a comprehensive exploration of UAV data collection methodologies for capturing heritage buildings in Montreal, addressing the limitations of existing data acquisition techniques, and the unique challenges posed by heritage structures. Our motivation stemmed from the scarcity of dedicated 3D datasets for heritage documentation and preservation, particularly considering Montreal's rich architectural heritage.

We commenced our endeavour by meticulously selecting heritage buildings with diverse architectural styles, considering factors such as historical significance and architectural uniqueness. Leveraging consumer-grade UAVs, namely the DJI Mini 2 and DJI Mini 3 Pro, equipped with advanced imaging capabilities, we devised an SOP for flight planning and control.

Throughout the data collection process, we encountered various challenges. Safety remained paramount throughout our UAV operations, with stringent adherence to regulations governing flight heights and proximity to restricted zones. We conducted thorough assessments of weather conditions and visitor logistics to mitigate potential risks to human life and property.

The resulting dataset, *Heritage3DMtl*, comprising 17 heritage buildings from mid-19th to early 20th-century structures in Montreal, stands as one of the largest real-world UAV datasets of its kind. Encompassing a range of architectural styles and cultural influences, it serves as a valuable resource for research and applications in structural engineering, cultural heritage preservation, and urban modelling.

Furthermore, our comparative analysis with aerial LiDAR and ground-level smartphone images highlighted the superiority of UAV data in capturing detailed architectural features and structural elements. Despite challenges such as occlusions and varying viewpoints, UAV data exhibited superior completeness and visual fidelity, showcasing its potential as a primary modality for heritage documentation and preservation efforts.

In conclusion, our study underscores the efficacy of UAV technology in generating multi-modal datasets for heritage buildings, offering unprecedented insights into Montreal's architectural legacy. By bridging the gap in 3D data availability, our work paves the way for enhanced documentation, preservation, and analysis of historical structures.

Table 3: Year of establishment and a brief description of the buildings captured

Sr. No.	**Building Name**	**Year Established**	**Architecture Style**	**Description**
1	Montreal City Hall	1878	Beaux-Arts	A National Historic Site of Canada, it is the first city hall to be constructed in the country.
2	Clock Tower	1922	Beaux-Arts	A Classified Federal Heritage Building whose construction was dedicated to commemorate the sailors who died during the first World War.
3	Saint Joseph's Oratory	1904	Italian Renaissance	A National Historic Site of Canada and is Canada's largest church, attracting millions of pilgrims and visitors every year.
4	Loyola Chapel	1933	Gothic Revival	A Roman Catholic church multi-faith community space located on Concordia University's Loyola campus.
5	Holy Ghost Ukrainian Catholic Church	1947	Kievan Rus	An onion-domed church that features Russian architecture.
6	Westmount City Hall	1922	Tudor Revival	A local government building, built in Neo-Tudor style is a reminiscent of Scottish castles.
7	St. George Antiochian Orthodox Church	1940	Predominantly Byzantine	A designated a National Historic Site of Canada as an important symbol of the history and traditions of the Syrian Orthodox community in Canada.
8	Maisonneuve Market	1912	Beaux-Arts	A public market that follows Beaux-Arts architectural style.
9	Maison Louis-Hippolyte Lafontaine	mid-1840s	Neoclassical Victorian	Once home to the first Prime Minister of the United Canadas, stands as a significant monument of Canadian political history and heritage.
10	Shaughnessy Mansion	1874	Second Empire style	An elegant house recognized as a National Historic Site for exemplifying the architectural style of its era and Montreal's greystone tradition.
11	St. Thomas More Church	1951	Modernist	A simple yet unique 5-facade church featuring steeple.
12	Resurrection Chapel	-	Victorian Gothic Revival	A chapel in Notre-Dame-des-Neiges cemetery, which is the largest cemetery in Canada; established in 1854, and has been recognized as a national historic site.
13	Murugan Temple	1983	Saivaite	This Hindu temple is the first Saivite temple in Quebec, and features the Saivaite Architectural design.
14	Hindu Mandir	1990	Hindu	This Hindu temple was built to serve the Indo-Canadian community in Montreal.
15	Gurdwara Guru Nanak Darbar, LaSalle	2001	Sikh	A Sikh temple that combines elements of historic Sikh design with Western postmodernist aesthetics. It stands at 172 feet high and is unique because it is one of only 12 worldwide.
16	Congregation Shaar Hashomayim	1922	Byzantine Revival	It is the oldest Ashkenazi synagogue and the largest traditional synagogue in Canada.
17	Islamic Center of Quebec - El Markaz Islami	1965	Islamic	It is the oldest mosque in Quebec and the second oldest in Canada.

Chapter 4

Experiments: Dataset Processing and Applications

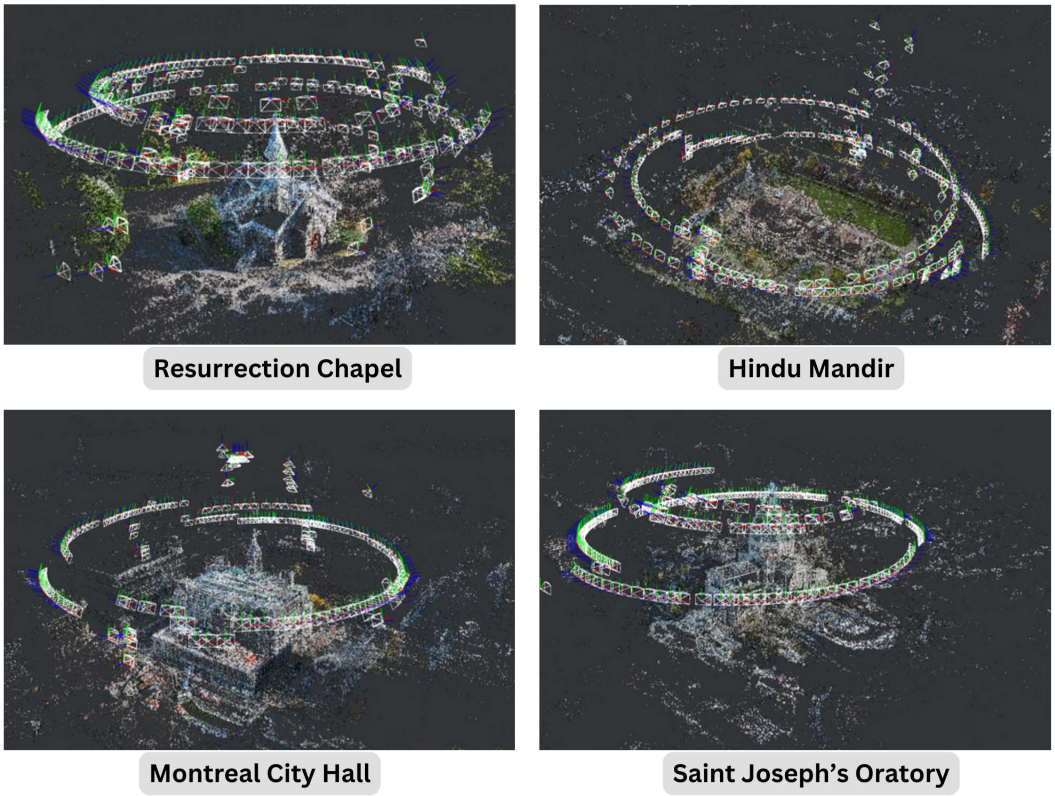
In this chapter, we discuss how we processed our collected data to generate multiple geometric representations using different 3D reconstruction techniques. We further demonstrate the compliance and utility of our dataset by generating LOD models. We also introduce a novel component to existing workflows for LOD3 generation using zero-shot capabilities of vision-language models.

4.1 SfM + MVS Reconstruction

Point clouds and meshes derived from UAVs or other data modalities play a central role in classical 3D building modelling [78–80]. Photogrammetry, combining SfM and MVS techniques, is a widely adopted method for generating these 3D representations.

In our study, we employ Meshroom to create point clouds and meshes for each building from the images captured in our dataset. Meshroom offers an intuitive node-based interface for managing the reconstruction process. Each node corresponds to a particular task, interconnected by edges.

Our reconstruction pipeline begins with importing the images and initiating the computation. The first node,*CameraInit*, initializes camera's intrinsic and image metadata. Next, the *FeatureExtraction* node detects feature points in the images, providing raw data for matching. Subsequently, the *ImageMatching* node identifies which images capture the same region, and the *FeatureMatching* node precisely aligns



Resurrection Chapel

Hindu Mandir

Montreal City Hall

Saint Joseph's Oratory

Figure 12: Examples of camera locations estimated by Meshroom

corresponding points across these pairs. Utilizing this data, the *StructureFromMotion* node estimates camera poses and reconstructs a sparse point cloud (see Fig. 12).

Subsequently, we export the camera pose data (intrinsics + extrinsics) in a JSON format. The *PrepareDenseScene* node then prepares the data for a denser point cloud creation, followed by the *DepthMap* node, which calculates depth information for each image. The *DepthMapFilter* node refines these depth maps, filtering out the noise and inconsistencies. The *Meshing* node generates a 3D mesh from these refined depth maps, which is later textured by the *Texturing* node to provide a realistic surface appearance.

The resulting point clouds and meshes often encompass more than just the target building. Therefore, we export them to Meshlab for further processing. Using the selection tool, we highlight the areas not belonging to the building structure and remove them. This step filters out noise, reduces the number of points, enhances clarity and reduces file size. This also ensures that the representations are optimized for subsequent applications.

4.2 Scene Representation

4.2.1 Background and Related Work

In Chapter 2, we discussed how NeRFs utilize ray marching techniques to determine the position along the viewing angle [25], enabling the reconstruction of 3D scenes. NeRF learns the scene from images and represents it as a function of 3D space. 3D Gaussian Splatting [26] employs a cloud of 3D Gaussians optimized through SGD for real-time, photorealistic volume rendering from any viewpoint. Recent research has investigated the effectiveness of NeRF and Gaussian Splatting for heritage 3D reconstruction, highlighting their effectiveness in describing material characteristics and reducing processing times compared to the traditional MVS techniques [71,81,82]. Both techniques are capable of handling challenging lighting conditions, occlusions, and complex surface textures, resulting in accurate and immersive visualizations. In our experiments, we aim to evaluate and compare the effectiveness of these techniques in accurately reconstructing heritage buildings. Our focus lies on assessing geometric fidelity by reconstructing mesh models and the quality of novel-view synthesis to identify the most effective methods for the digital preservation and visualization of cultural heritage sites.

4.2.2 Implementation Details

Nerstudio [83] is a modular PyTorch framework that provides a simple API that allows for a simplified end-to-end process of creating, training, and testing NeRFs. The framework's modular design supports real-time visualization tools, streamlined pipelines for importing real-world data, and tools for exporting to video, point cloud, and mesh formats. Nerfacto is a method implemented by Nerstudio, which leverages the modular design of the framework to combine components from recent NeRF implementations and achieve a balance of speed and quality. This approach aims to achieve a balance between speed and quality in NeRF models, maintaining flexibility for future modifications. The nerfacto method incorporates optimized camera views, ray generation and sampling strategies, and scene contraction techniques to efficiently and effectively process and render 3D scenes. In our experiments, we use nerfacto to generate novel views and export meshes. Because the 3D Gaussian splats are not necessarily aligned with the surface, the original Gaussian Splatting method does not allow for a good mesh reconstruction. Therefore, for mesh generation, we use a variation of the original method: Surface-Aligned Gaussian Splatting

(SuGaR) [84], which enables a better mesh reconstruction. The key to better mesh reconstruction in SuGaR is the introduction of a regularization term designed to align the splats/gaussians closely with the surface geometry of the scene. This regularization term encourages the Gaussians to align with and accurately capture the geometry of the scene, thus facilitating a better mesh reconstruction process.

4.2.3 Results and Discussion

We compare the three approaches for the fidelity of the reconstructed geometry as well as novel-view synthesis applications. The qualitative results comparing mesh reconstructions from our experiments are illustrated in Fig. 13. We observe that SuGaR is able to preserve surface textures and capture fine-scale geometric features with higher fidelity better than nerfacto. Overall, photogrammetry still excels in geometric accuracy and detailed surface structure. In terms of novel view synthesis, both Gaussian Splatting and nerfacto exhibited the ability to create highly realistic scenes with impressive visual coherence across different viewpoints. However, nerfacto was observed to exhibit a higher tendency to produce artifacts and blurriness, particularly in areas with sparse or inconsistent image coverage. This suggests that NeRF encounters challenges in interpolating scene structure and lighting in areas with

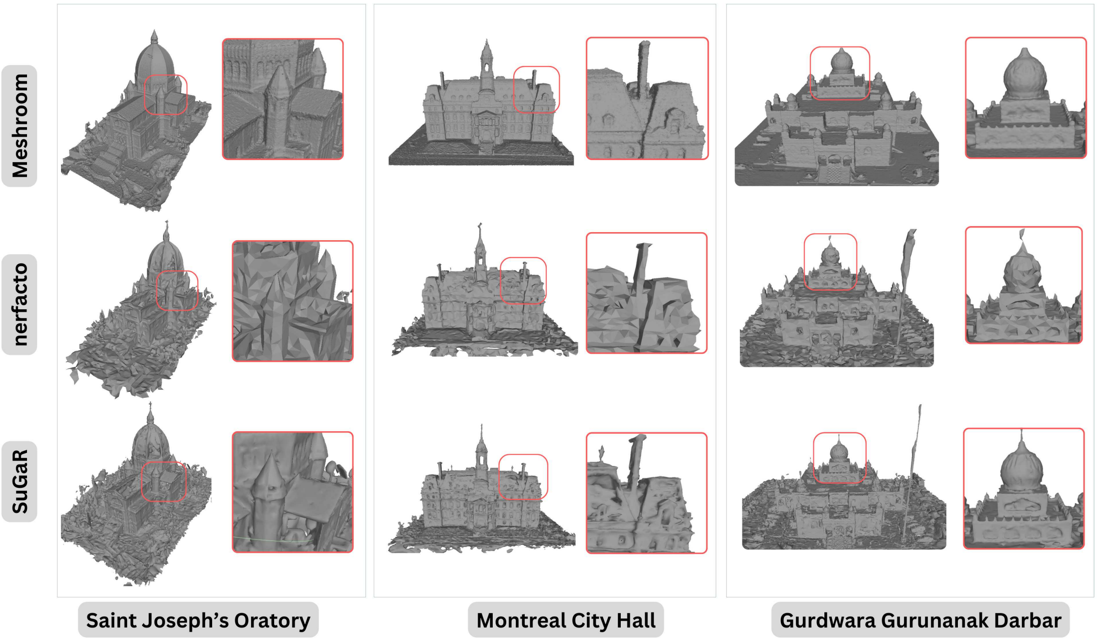


Figure 13: Meshes generated from Meshroom (MVS), nerfacto and Surface-Aligned Guassian Splatting (SuGaR)

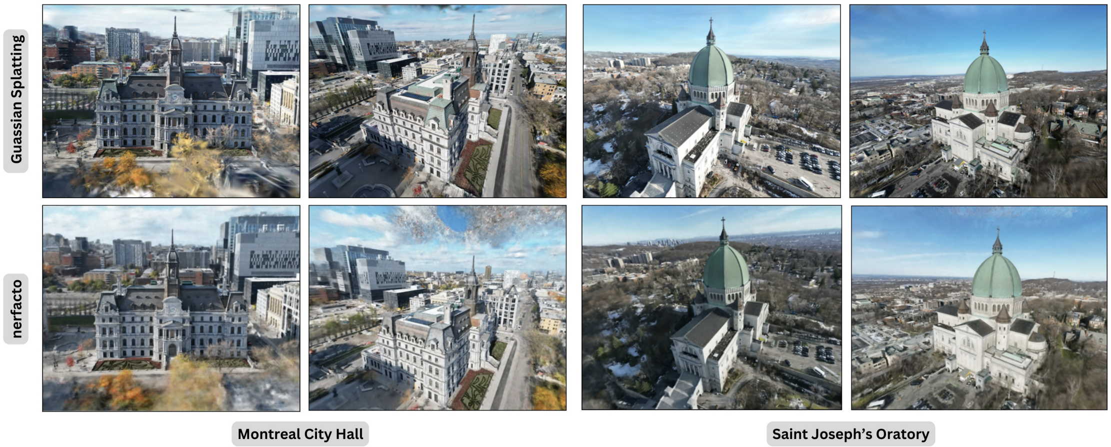


Figure 14: Novel views generated by 3D Gaussian Splatting and NeRF (nerfacto)

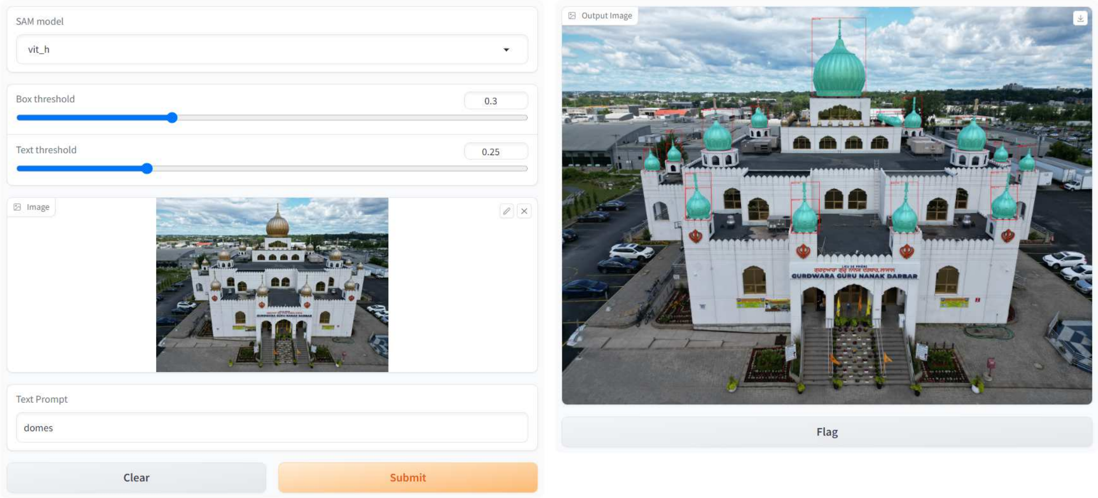
insufficient data, resulting in diminished visual quality. On the other hand, Gaussian Splatting results show higher sharpness and realistic shadows, leading to more photorealistic results. Representative samples that illustrate these differences in synthesis quality are displayed in Fig. 14, which showcase the comparative renderings from similar novel viewpoints. Since Gaussian Splatting achieves high-quality renderings at high frame rates, it holds significant potential for applications such as real-time visualization and virtual tourism. This technique could be instrumental in delivering immersive experiences of heritage sites, enhancing user engagement.

4.3 Level of Detail 3 (LOD3) Modelling

4.3.1 Background and Related Work

BIM finds extensive use in architecture, structural engineering, and construction industries for 3D digital representations of buildings. However, BIM modelling frameworks often require significant time and effort, leading to various errors and necessitating manual inspection or rule set development [85].

In contrast, CityGML 3.0 defines four levels of abstraction, ranging from LOD 0-3, as discussed in Chapter 2, providing a standardized approach to represent buildings and city-scale data. It facilitates the classification of computational building models for various applications, from urban planning to structural analysis. While detailed models derived from BIM or from techniques like MVS are essential for certain applications, others benefit from the simplified geometric primitives of LOD models.



SAM model

vit_h

Box threshold 0.3

Text threshold 0.25

Text Prompt

domes

Figure 15: Result of Grounding DINO + SAM with text prompt: domes

However, existing methodologies for automated 3D reconstruction often face challenges at higher LODs, lacking robustness and semantic understanding or requiring manual intervention [86]. The choice of LOD depends on the specific requirements of the application or use case. For instance, urban planners may prefer LOD2 models to visualize and analyze the general layout of a city.

In the context of heritage buildings, LOD3 models hold significance as they capture building semantics, enable abstract representation, and facilitate documentation, structural analysis, and integration into digital urban models [87, 88].

4.3.2 Implementation Details

In our extended workflow for LOD3 construction, we enhance the pipeline initially proposed in [86], which aims to generate LOD3 models of stone masonry buildings automatically.

We introduce the use of zero-shot approach that integrates the SAM [43] and Grounding DINO [44] for the detection and generation of 2D masks of architectural features in facade images, which are further used for automated generation of LOD3 models. Integration of SAM and Grounding DINO allows for detecting and segmenting objects or regions in images based on arbitrary text inputs. This approach was proposed in Grounded SAM [89]. It is effective for a wide array of tasks, across both common and long-tail object categories. It first uses Grounding DINO to identify objects with text and then applies SAM for mask generation. As discussed

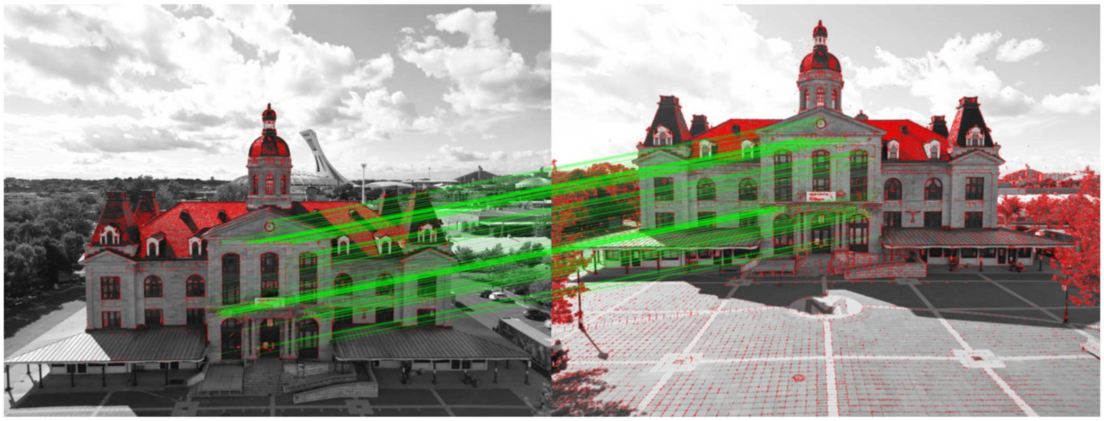


Figure 16: Matching filtered key points across two views

earlier, SAM is a foundation model for image segmentation that enables zero-shot transfer to new image distributions and tasks by interpreting prompts that specify what to segment within an image. Grounding DINO is an open-set object detector that employs a transformer-based architecture and enables the detection of arbitrary objects by processing text input. The combination of these two models significantly streamlines our workflow, does semantic feature extraction enabling us to identify objects of interest in the image like, like windows, doors, stairs, domes, pillars, frames, etc., as depicted in Fig. 15 and eliminates the need for manual image annotations, which are often expensive and labour-intensive to create and also does not require further training of the network.

To construct a LOD2 model, we utilize point clouds generated from Meshroom and employ the Polyfit framework [5]. This method starts with the extraction of planar primitives from the point cloud using RANSAC. These are further refined through merging and fitting new planes, coupled with computing intersections to propose candidate faces. An optimal subset of faces is chosen through binary linear programming optimization, ensuring the resulting polygonal surface model is both manifold and watertight. We use the Gurobi optimizer in our experiments. This resulting reconstruction is the LOD2 model, a lightweight polygonal surface model that preserves sharp features and is resilient against noise and outliers.

Further, we generate segmented masks for doors and windows in all captured facade images of the building using SAM + Grounding DINO. [90]. To further elevate the LOD2 models to LOD3, we integrate the boundaries of the openings (windows, doors) detected by triangulating them. The triangulation is achieved through a series of steps starting with the detection and matching of key points between two views using the SIFT algorithm [91]. The detected key points undergo a filtering process, ensuring that only those on the same plane (the building facade) are considered for further processing. The correspondence between these filtered points across views is established through their SIFT descriptors, further which homography matrix H is computed using the Direct Linear Transformation (DLT) algorithm [38] which maps points $\mathbf{x}_i$ in one plane to corresponding points $\mathbf{x}'_i$ in another plane. This is depicted in Fig. 16. The correspondences relate to H as Eq. (3).

$$x'_i = Hx_i \tag{3}$$

Subsequently, the process of triangulating these point correspondences to 3D space is based on epipolar geometry, using the camera projection matrices derived from SfM camera poses, as discussed in Chapter 2.

Finally, these openings are geometrically subtracted from the LOD2 models to obtain LOD3 models using implementation from FreeCAD library [92]. Our entire workflow for generating LOD3 from UAV images is depicted in Fig. 17.

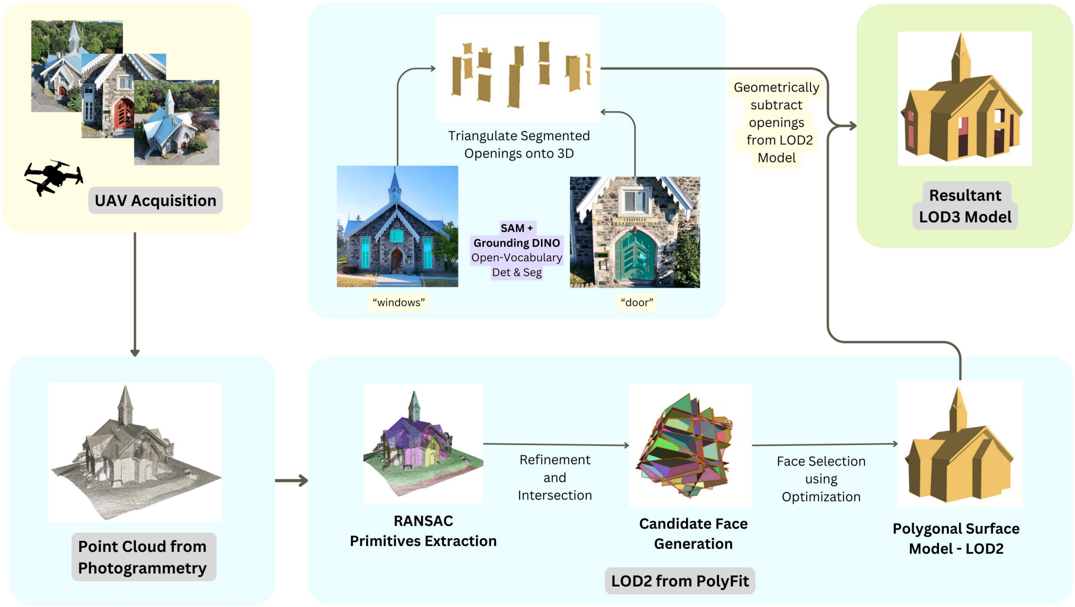

UAV Acquisition
Triangulate Segmented Openings onto 3D
SAM + Grounding DINO Open-Vocabulary Det & Seg
"windows"
"door"
Geometrically subtract openings from LOD2 Model
Resultant LOD3 Model
Point Cloud from Photogrammetry
RANSAC Primitives Extraction
Refinement and Intersection
Candidate Face Generation
Face Selection using Optimization
Polygonal Surface Model - LOD2
LOD2 from PolyFit

Figure 17: Overview of the workflow for LOD3 generation

4.3.3 Results and Discussion

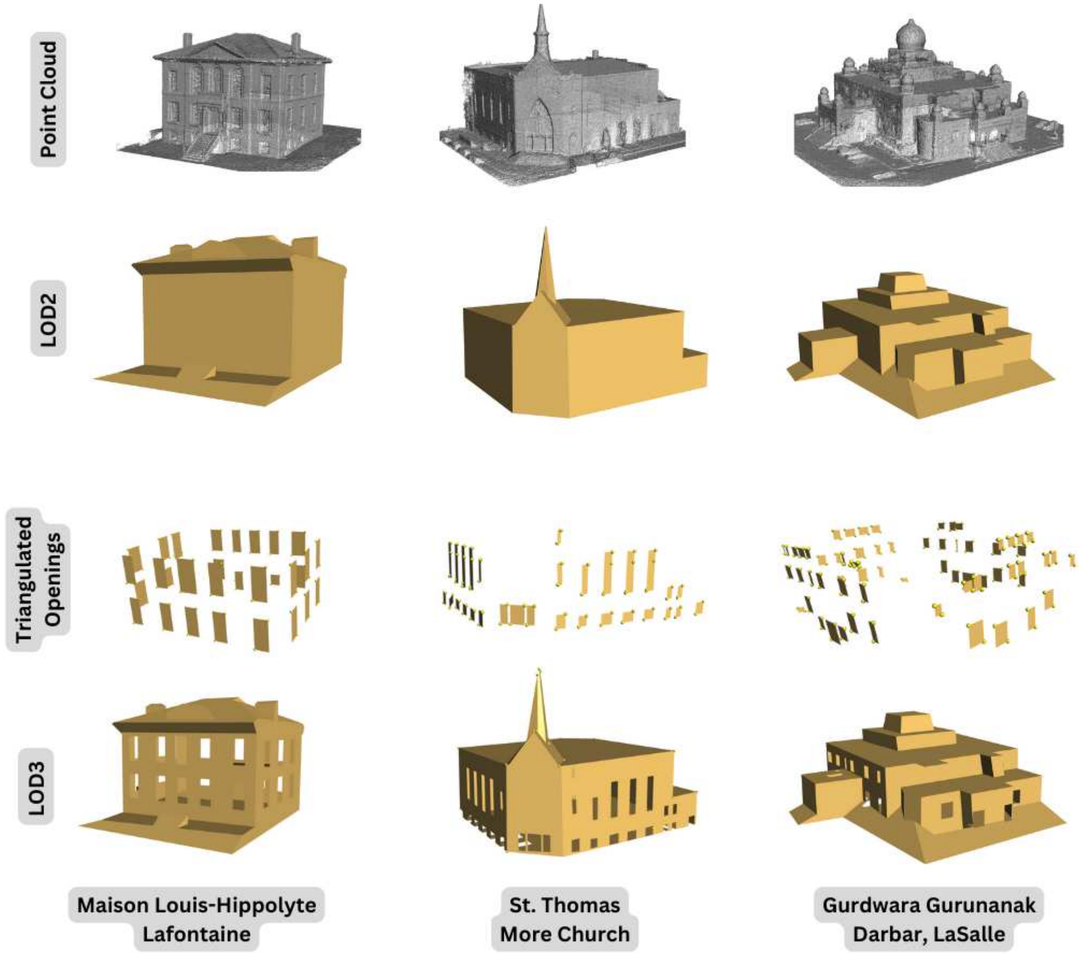

Point Cloud
LOD2
Triangulated Openings
LOD3
Maison Louis-Hippolyte Lafontaine
St. Thomas More Church
Gurdwara Gurunanak Darbar, LaSalle

Figure 18: Sample results of generated LOD2 and LOD3 models on our dataset.

Our approach in utilizing open-vocabulary detection and segmentation works well for extracting building semantics like different architectural elements heavily present in heritage buildings. We have successfully demonstrated its use to elevate LOD2 models to LOD3 models in an automated manner, significantly streamlining existing pipelines. Visual comparison of MVS mesh, LOD2 and LOD3 results from our experiments for three buildings in our dataset are depicted in Fig. 18. Statistics of some results are given in Tab. 2.

Table 4: Statistics of reconstructed LOD models

Building	**# Vertices in input Point Cloud**	**# Faces in LOD2**	**# Faces in LOD3**
Maison Louis-Hippolyte Lafontaine	972 K	1092	1366
St. Thomas More Church	827 K	462	322
Gurdwara Guru Nanak Darbar	1.06 M	1180	1579
Westmount City Hall	1.01 M	364	861
Resurrection Chapel	656 K	2342	1297

4.4 Conclusion

In this chapter, we have explored various methods and techniques for reconstructing and modelling heritage buildings at different LODs. By leveraging advancements in computer vision, machine learning, and geometric modelling, we aimed to address the challenges associated with capturing, representing, and preserving cultural sites.

First, we discussed scene representation techniques, including SfM, MVS, and recent advancements such as NeRFs and Gaussian Splatting. Moving on to LOD modelling, we presented an extended workflow that integrates semantic segmentation models with geometric reconstruction techniques to automatically generate detailed representations of heritage buildings. By integrating SAM and Grounding DINO for object detection and segmentation, we showcased the ability to derive building semantics and upgrade LOD2 models to LOD3 models automatically.

Our results indicate that the proposed approaches offer significant advancements in the field of heritage building reconstruction. Photogrammetry techniques excel in geometric accuracy and surface structure detailing, while NeRFs and Gaussian Splatting show promise in novel view synthesis applications. Furthermore, integration of semantic segmentation models streamlines the process of generating detailed LOD3 models, providing valuable insights for documentation, preservation, and analysis of cultural heritage sites.

In conclusion, the methodologies presented in this chapter contribute towards the advancement of digital preservation and visualization of heritage buildings. Combining state-of-the-art reconstruction techniques with semantic understanding, we pave the way for more efficient and accurate representations of cultural heritage, fostering greater accessibility and appreciation of our architectural legacy.

Chapter 5

Conclusions and Future Work

5.1 Introduction

Preserving and documenting cultural heritage is a multifaceted endeavour that requires the integration of diverse technologies and methodologies. Over the years, advancements in fields such as computer vision, remote sensing, and UAV technology have reshaped our approach to heritage preservation. These technologies present unprecedented opportunities to capture, analyze, and interpret heritage sites in three dimensions, providing valuable insights into our shared history and cultural identity.

In recent years, the adoption of UAVs has emerged as a game-changer in heritage documentation. Equipped with high-resolution cameras and LiDAR sensors, UAVs enable researchers to capture detailed aerial imagery and generate precise 3D models of heritage structures. This aerial perspective not only facilitates the identification of structural vulnerabilities and preservation needs but also allows for the creation of immersive virtual experiences that engage audiences in exploration and appreciation of cultural heritage.

Furthermore, the integration of advanced reconstruction techniques, such as NeRF and Gaussian Splatting, has revolutionized the way we visualize and interact with heritage sites in digital environments. These techniques enable the generation of photorealistic renderings as well as interactive 3D models, enhancing our ability to study and interpret heritage buildings with unprecedented fidelity and detail. This fusion of cutting-edge technologies empowers researchers and preservationists to delve deeper into the architectural intricacies and historical significance of heritage sites.

In this thesis, we have embarked on a journey to explore the intersection of UAV-based data collection and advanced reconstruction methodologies for heritage preservation. Through the meticulous documentation of 17 heritage buildings in Montreal, Canada, using low-cost UAVs, we have created the *Heritage3DMtl* dataset—a comprehensive resource for researchers and practitioners in the field of heritage studies. By leveraging UAV-derived data and state-of-the-art reconstruction techniques, we have demonstrated the transformative potential of these technologies in the preservation and interpretation of cultural heritage.

5.2 Contributions

The research presented in this thesis contributes significantly to the advancement of 3D heritage reconstruction and preservation through the following key contributions:

1. **Heritage3DMtl Dataset:** The *Heritage3DMtl* dataset democratizes access to high-quality 3D data for heritage buildings. By capturing detailed geometric information and visual representations of 17 heritage buildings in Montreal, Canada, using UAVs, this dataset provides a valuable resource for researchers and practitioners in the fields of computer vision, remote sensing, and heritage studies.

2. **Low-Cost Capture Process:** The development of an effective capture process for detailed structures utilizing UAV-based data collection underscores the cost-effectiveness and scalability of this approach. By leveraging consumer-grade UAV platforms equipped with cameras, researchers can gather data with minimal upfront investment, facilitating large-scale data acquisition for heritage reconstruction projects.

3. **Advancements in Reconstruction Techniques:** The exploration and demonstration of state-of-the-art reconstruction techniques, including NeRF and Gaussian Splatting, highlight their efficacy in generating realistic and detailed renderings of heritage buildings. Additionally, the investigation of LOD modelling using open vocabulary zero-shot detection and segmentation showcases the potential for semantic understanding to enhance the interpretability and utility of reconstructed models.

5.3 Future Work

While this research represents a significant step forward in 3D heritage reconstruction and preservation, several avenues for future work warrant exploration:

1. **Integration of Semantic Segmentation:** We aim to explore the integration of semantic segmentation techniques to enhance LOD3 models beyond doors and window openings. By incorporating intricate surface structures such as walls, roofs, and window/door frames, we can create more comprehensive and detailed 3D models of heritage buildings, capturing their architectural richness and complexity in greater detail.

2. **Multi-Modal Integration:** Exploring the integration of multi-modal data sources, such as UAV imagery, LiDAR, and ground-based photographs, could enhance the accuracy and completeness of reconstructed models.

3. **Real-Time Visualization:** Investigating real-time visualization techniques for heritage preservation and education purposes could enable immersive and interactive experiences for stakeholders and the public.

In conclusion, the findings presented in this thesis underscore the transformative potential of UAV-based data collection and advanced reconstruction techniques in the domain of 3D heritage reconstruction and preservation. By leveraging these technologies responsibly and collaboratively, researchers can contribute to the digital documentation, preservation, and interpretation of our shared cultural heritage for future generations.

Appendix

Paper Submission to ECCV 2024: This work has resulted in a research paper that has been submitted to the European Conference on Computer Vision (ECCV) 2024 and is currently under review. Below is the abstract of the submitted paper:

Abstract: We present an extensive real-world, multi-view UAV dataset of 17 architecturally diverse heritage buildings in Montreal, along with estimated camera poses, point clouds, and dense triangle meshes reconstructed using open-source photogrammetry tools. This dataset not only includes Level of Detail 2 (LOD2) and Level of Detail 3 (LOD3) models in alignment with the CityGML standard but is also openly available to foster further research in 3D heritage reconstruction, structural analysis, and cultural heritage preservation, among other areas. Our discussion emphasizes the efficient and effective use of low-cost consumer-grade UAVs and the meticulous capture process employed. After careful consideration, we opted for UAVs over available LiDAR technology or the more accessible smartphone cameras due to the comprehensive reach that UAVs can provide. This choice enabled us to create detailed 3D geometric representations covering regions and details not easily achievable with other modalities. Furthermore, as a demonstration of its applicability, we reconstruct LOD3 models using our dataset. We also test the dataset with Neural Radiance Fields (NeRFs) and 3D Gaussian Splatting for novel view generation, showcasing its versatility. Compared to many other heritage datasets, which are often quite small and cover only one or two monuments, our dataset stands out for its extensive scope, diversity, and detail.

References

[1] P. Moulon, P. Monasse, and R. Marlet, "Adaptive structure from motion with a contrario model estimation," in *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part IV 11*, pp. 257–270, Springer, 2013. vii, 8

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. vii, 10

[3] M. S. M. Sajjadi, "A comprehensive overview of gaussian splatting," *Towards Data Science*, 2022. vii, 11

[4] "Ogc city geography markup language (citygml) 3.0 conceptual model users guide," 09 2021. Accessed: 2024-03-05. vii, 12

[5] L. Nan and P. Wonka, "Polyfit: Polygonal surface reconstruction from point clouds," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2353–2361, 2017. vii, 13, 36

[6] P. McEnroe, S. Wang, and M. Liyanage, "A survey on the convergence of edge computing and ai for uavs: Opportunities and challenges," *Ieee Internet of Things Journal*, 2022. 1

[7] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018. 1

[8] H. Obanawa, Y. Hayakawa, and C. Gomez, "3D Modelling of Inaccessible Areas using UAV-based Aerial Photography and Structure from Motion," in *EGU*

General Assembly Conference Abstracts, EGU General Assembly Conference Abstracts, p. 5063, May 2014. 1

[9] A. Abdallah, M. Z. Ali, J. Mišić, and V. B. Mišić, “Efficient security scheme for disaster surveillance uav communication networks,” *Information*, vol. 10, p. 43, 2019. 1

[10] D. Dominici, M. Alicandro, and V. Massimi, “Uav photogrammetry in the post-earthquake scenario: case studies in l’aquila,” *Geomatics, Natural Hazards and Risk*, vol. 8, pp. 87–103, 2016. 1

[11] Y. Liu, X. Nie, J. Fan, and X. Liu, “Image-based crack assessment of bridge piers using unmanned aerial vehicles and three-dimensional scene reconstruction,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 35, pp. 511–529, 2019. 1

[12] F. Ioli, A. Pinto, and L. Pinto, “Uav photogrammetry for metric evaluation of concrete bridge cracks,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B2-2022, pp. 1025–1032, 2022. 1

[13] Y. Lin, Y. Cheng, T. Zhou, R. Ravi, S. M. Hasheminasab, J. E. Flatt, C. D. Troy, and A. Habib, “Evaluation of uav lidar for mapping coastal environments,” *Remote Sensing*, vol. 11, p. 2893, 2019. 1

[14] M. Jaud, P. Letortu, C. Théry, P. Grandjean, S. Costa, O. Maquaire, R. Davidson, and N. L. Dantec, “Uav survey of a coastal cliff face – selection of the best imaging angle,” *Measurement*, vol. 139, pp. 10–20, 2019. 1

[15] G. Brunner, B. Szebedy, S. Tanner, and R. Wattenhofer, “The urban last mile problem: autonomous drone delivery to your balcony,” *2019 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2019. 1

[16] S. Harwin and A. Lucieer, “Assessing the accuracy of georeferenced point clouds produced via multi-view stereopsis from unmanned aerial vehicle (uav) imagery,” *Remote Sensing*, vol. 4, pp. 1573–1599, 2012. 1

[17] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, “Depth-supervised nerf: Fewer views and faster training for free,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12882–12891, 2022. 1

[18] F. Condorelli, F. Rinaudo, F. Salvadore, and S. Tagliaventi, "A comparison between 3d reconstruction using nerf neural networks and mvs algorithms on cultural heritage images," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B2-2021, pp. 565–570, 2021. 1

[19] C. Lu, F. Yin, X. Chen, W. Liu, T. Chen, G. Yu, and J. Fan, "A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7557–7567, 2023. 1, 17

[20] Y. Li, L. Jiang, L. Xu, Y. Xiangli, Z. Wang, D. Lin, and B. Dai, "Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3205–3215, 2023. 1, 17

[21] V. Bouzas, H. Ledoux, and L. Nan, "Structure-aware building mesh polygonization," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 432–442, 2020. 1

[22] E. Colucci, V. De Ruvo, A. Lingua, F. Matrone, and G. Rizzo, "Hbim-gis integration: From ifc to citygml standard for damaged cultural heritage in a multiscale 3d gis," *Applied Sciences*, vol. 10, no. 4, p. 1356, 2020. 1

[23] T. Shinohara, L. YongHe, M. Sakamoto, and T. Satoh, "Building cad model reconstruction from point clouds via instance segmentation, signed distance function, and graph cut," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1735–1744, 2023. 1

[24] T. Kutzner, K. Chaturvedi, and T. Kolbe, "Citygml 3.0: New functions open up new applications," *PFG – Journal of Photogrammetry Remote Sensing and Geoinformation Science*, vol. 88, 02 2020. 2

[25] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "Nerf–: Neural radiance fields without known camera parameters," 2022. 4, 32

[26] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023. 5, 10, 32

[27] R. Yaagoubi and Y. Miky, “Developing a combined light detecting and ranging (lidar) and building information modeling (bim) approach for documentation and deformation assessment of historical buildings,” *MATEC Web of Conferences*, vol. 149, p. 02011, 2018. 6

[28] D. Che, Z. Li, Y. Liu, R. Zhong, and B. Ma, “A new method of achieving single three-dimensional building model automatically based on oblique photography data,” *Mathematical Problems in Engineering*, vol. 2021, pp. 1–12, 2021. 6

[29] K. K. Mwangangi, P. O. Mc’Okeyo, S. O. Elberink, and F. Nex, “Exploring the potentials of uav photogrammetric point clouds in façade detection and 3d reconstruction of buildings,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B2-2022, pp. 433–440, 2022. 6, 17

[30] M. Li and L. Nan, “Feature-preserving 3d mesh simplification for urban buildings,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 135–150, 2021. 6

[31] C. Zhao, C. Zhang, Y. Yan, and N. Su, “A 3d reconstruction framework of buildings using single off-nadir satellite image,” *Remote Sensing*, vol. 13, no. 21, p. 4434, 2021. 6

[32] G. Park, C. Kim, M. Lee, and C. Choi, “Building geometry simplification for improving mesh quality of numerical analysis model,” *Applied Sciences*, vol. 10, no. 16, p. 5425, 2020. 6

[33] “Photogrammetry pipeline from alicevision,” vol. 00. `https://alicevision.org`. 6

[34] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[35] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, “OpenMVG: Open multiple view geometry,” in *International Workshop on Reproducible Research in Pattern Recognition*, pp. 60–74, Springer, 2016. 6

[36] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, pp. 91–110, 2004. 7

[37] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Ep n p: An accurate o (n) solution to the p n p problem," *International journal of computer vision*, vol. 81, pp. 155–166, 2009. 8

[38] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003. 9, 37

[39] K. Chaidas, G. Tataris, and N. Soulakellis, "Seismic damage semantics on post-earthquake lod3 building models generated by uas," *ISPRS International Journal of Geo-Information*, vol. 10, p. 345, 2021. 12

[40] E. Colucci, F. Noardo, F. Matrone, A. T. Spano, and A. M. Lingua, "High-level-of-detail semantic 3d gis for risk and damage representation of architectural heritage," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-4, pp. 107–114, 2018. 13

[41] E. Colucci, V. D. Ruvo, A. M. Lingua, F. Matrone, and G. Rizzo, "Hbim-gis integration: from ifc to citygml standard for damaged cultural heritage in a multiscale 3d gis," *Applied Sciences*, vol. 10, p. 1356, 2020. 13

[42] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," in *Computer graphics forum*, vol. 26, pp. 214–226, Wiley Online Library, 2007. 13

[43] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. 15, 35

[44] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023. 35

[45] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arxiv 2022," *arXiv preprint arXiv:2203.03605*, vol. 5, 2022. 15

[46] F. Tarsha-Kurdi, T. Landes, P. Grussenmeyer, and M. Koehl, "Model-driven and data-driven approaches using lidar data: Analysis and comparison," in *ISPRS workshop, photogrammetric image analysis (PIA07)*, pp. 87–92, 2007. 16

[47] F. Rottensteiner and J. Jansa, "Automatic extraction of buildings from lidar data and aerial images," *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34, no. 4, pp. 569–574, 2002. 16

[48] M. Kada and L. McKinley, "3d building reconstruction from lidar based on a cell decomposition approach," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, no. Part 3, p. W4, 2009. 16

[49] R. Wang, S. Huang, and H. Yang, "Building3d: A urban-scale dataset and benchmarks for learning roof structures from point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20076–20086, 2023. 16

[50] S. Daftry, C. Hoppe, and H. Bischof, "Building with drones: Accurate 3d facade reconstruction using mavs," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3487–3494, IEEE, 2015. 16

[51] D. Lapandic, J. Velagic, and H. Balta, "Framework for automated reconstruction of 3d model from multiple 2d aerial images," in *2017 International Symposium ELMAR*, pp. 173–176, IEEE, 2017. 16

[52] A. Filatov, M. Zaslavskiy, and K. Krinkin, "Multi-drone 3d building reconstruction method," *Mathematics*, vol. 9, no. 23, p. 3033, 2021. 16

[53] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12922–12931, 2022. 16, 17

[54] A. Kyriou, K. Nikolakopoulos, and I. Koukouvelas, "Synergistic use of uav and tls data for precise rockfall monitoring over a hanging monastery.," in *Earth Resources and Environmental Remote Sensing/GIS Applications XIII*, vol. 12268, pp. 34–45, SPIE, 2022. 16, 17

[55] Z. Xu, L. Wu, Y. Shen, F. Li, Q. Wang, and R. Wang, "Tridimensional reconstruction applied to cultural heritage with the use of camera-equipped uav and terrestrial laser scanner," *Remote sensing*, vol. 6, no. 11, pp. 10413–10434, 2014. 17

[56] Y. H. Jo and S. Hong, "Three-dimensional digital documentation of cultural

heritage site based on the convergence of terrestrial laser scanning and unmanned aerial vehicle photogrammetry," *ISPRS International Journal of Geo-Information*, vol. 8, no. 2, p. 53, 2019. 17

[57] R. Andaru, B. Cahyono, G. Riyadi, G. Ramadhan, S. Tuntas, *et al.*, "The combination of terrestrial lidar and uav photogrammetry for interactive architectural heritage visualization using unity 3d game engine," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 39–44, 2019. 17

[58] P. Klapa and P. Gawronek, "Synergy of geospatial data from tls and uav for heritage building information modeling (hbim)," *Remote Sensing*, vol. 15, no. 1, p. 128, 2022. 17

[59] T. Luhmann, M. Chizhova, and D. Gorkovchuk, "Fusion of uav and terrestrial photogrammetry with laser scanning for 3d reconstruction of historic churches in georgia," *Drones*, vol. 4, no. 3, p. 53, 2020. 17

[60] K. Themistocleous, M. Ioannides, A. Agapiou, and D. G. Hadjimitsis, "The methodology of documenting cultural heritage sites using photogrammetry, uav, and 3d printing techniques: the case study of asinou church in cyprus," in *Third International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2015)*, vol. 9535, pp. 312–318, SPIE, 2015. 17

[61] F. Samadzadegan, F. Dadrass Javan, and M. Zeynalpoor Asl, "Architectural heritage 3d modelling using unmanned aerial vehicles multi-view imaging," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, pp. 1395–1402, 2023. 17

[62] E. Karachaliou, E. Georgiou, D. Psaltis, and E. Stylianidis, "Uav for mapping historic buildings: From 3d modelling to bim," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 397–402, 2019. 17

[63] A. Murtiyoso and P. Grussenmeyer, "Documentation of heritage buildings using close-range uav images: dense matching issues, comparison and case studies," *The Photogrammetric Record*, vol. 32, no. 159, pp. 206–229, 2017. 17

[64] G. Rizzoli, F. Barbato, M. Caligiuri, and P. Zanuttigh, "Syndrone-multi-modal uav dataset for urban scenarios," in *Proceedings of the IEEE/CVF International*

Conference on Computer Vision, pp. 2210–2220, 2023. 17

[65] S. Yan, X. Cheng, Y. Liu, J. Zhu, R. Wu, Y. Liu, and M. Zhang, "Render-and-compare: Cross-view 6-dof localization from noisy prior," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2171–2176, 2023. 17

[66] L. Lin, Y. Liu, Y. Hu, X. Yan, K. Xie, and H. Huang, "Capturing, reconstructing, and simulating: the urbanscene3d dataset," in *European Conference on Computer Vision*, pp. 93–109, Springer, 2022. 17

[67] X. Zhang, X. Cao, A. Yu, W. Yu, Z. Li, and Y. Quan, "Uavstereo: A multiple resolution dataset for stereo matching in uav scenarios," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 2942–2953, 2023. 17

[68] M. Fonder and M. Van Droogenbroeck, "Mid-air: A multi-modal dataset for extremely low altitude drone flights," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019. 17

[69] M. Chen, Q. Hu, Z. Yu, H. Thomas, A. Feng, Y. Hou, K. McCullough, F. Ren, and L. Soibelman, "Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset," *arXiv preprint arXiv:2203.09065*, 2022. 17

[70] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham, "Sensaturban: Learning semantics from urban-scale photogrammetric point clouds," *International Journal of Computer Vision*, vol. 130, no. 2, pp. 316–343, 2022. 17

[71] G. Mazzacca, A. Karami, S. Rigon, E. Farella, P. Trybala, and F. Remondino, "Nerf for heritage 3d reconstruction," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, pp. 1051–1058, 2023. 20, 32

[72] J. Ryding, E. Williams, M. J. Smith, and M. P. Eichhorn, "Assessing handheld mobile laser scanners for forest surveys," *Remote Sensing*, vol. 7, pp. 1095–1111, 2015. 20

[73] L. Terryn, K. Calders, M. Disney, N. Origo, Y. Malhi, G. Newnham,

P. Raumonen, M. kerblom, and H. Verbeeck, “Tree species classification using structural features derived from terrestrial laser scanning,” *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 170–181, 2020. 20

[74] Y. Alshawabkeh, A. Baik, and Y. Miky, “Integration of laser scanner and photogrammetry for heritage bim enhancement,” *Isprs International Journal of Geo-Information*, vol. 10, p. 316, 2021. 20

[75] S. Daftry, C. Hoppe, and H. Bischof, “Building with drones: Accurate 3d facade reconstruction using mavs,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3487–3494, 2015. 22

[76] E. Seifert, S. Seifert, H. Vogt, D. M. Drew, J. v. Aardt, A. Kunneke, and T. Seifert, “Influence of drone altitude, image overlap, and optical sensor resolution on multi-view reconstruction of forest images,” *Remote Sensing*, vol. 11, p. 1252, 2019. 22

[77] J. Frey, K. Kovach, S. Stemmler, and B. Koch, “Uav photogrammetry of forests as a vulnerable process. a sensitivity analysis for a structure from motion rgb-image pipeline,” *Remote Sensing*, vol. 10, p. 912, 2018. 22

[78] F. Remondino and S. El-Hakim, “Image-based 3d modelling: a review,” *The photogrammetric record*, vol. 21, no. 115, pp. 269–291, 2006. 30

[79] F. Alidoost and H. Arefi, “An image-based technique for 3d building reconstruction using multi-view uav images,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, pp. 43–46, 2015. 30

[80] Y. Liang, X. Fan, Y. Yang, D. Li, and T. Cui, “Oblique view selection for efficient and accurate building reconstruction in rural areas using large-scale uav images,” *Drones*, vol. 6, no. 7, p. 175, 2022. 30

[81] F. Comte, A. Pamart, K. Réby, and L. De Luca, “Strategies and experiments for massive 3d digitalization of the remains after the notre dame de paris’ fire,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-2/W4-2024, pp. 127–134, 2024. 32

[82] V. Croce, D. Billi, G. Caroti, A. Piemonte, L. De Luca, and P. Véron, “Comparative assessment of nerf and photogrammetry in digital heritage:

Impact of varying image conditions on 3d reconstruction," 2023. 32

[83] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, *et al.*, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–12, 2023. 32

[84] A. Guédon and V. Lepetit, "Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering," *arXiv preprint arXiv:2311.12775*, 2023. 33

[85] H. Liu, J. Cheng, V. Gan, and S. Zhou, "Automatic bim model auditing for quantity take-off using knowledge graph techniques," in *IOP Conference Series: Earth and Environmental Science*, vol. 1101, p. 092031, IOP Publishing, 2022. 34

[86] B. Pantoja-Rosero, R. Achanta, M. Kozinski, P. Fua, F. Perez-Cruz, and K. Beyer, "Generating lod3 building models from structure-from-motion and semantic segmentation," *Automation in Construction*, vol. 141, p. 104430, 2022. 35

[87] B. Ergun, C. Sahin, and F. Bilucan, "Level of detail (lod) geometric analysis of relief mapping employing 3d modeling via uav images in cultural heritage studies," *Heritage Science*, vol. 11, no. 1, p. 194, 2023. 35

[88] E. Colucci, F. Noardo, F. Matrone, A. Spanò, and A. Lingua, "High-level-of-detail semantic 3d gis for risk and damage representation of architectural heritage," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 107–114, 2018. 35

[89] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024. 35

[90] L. Medeiros, "lang-segment-anything: Language-agnostic text segmentation for any language." `https://github.com/luca-medeiros/lang-segment-anything`, Year. 36

[91] G. Lowe, "Sift-the scale invariant feature transform," *Int. J*, vol. 2, no. 91-110, p. 2, 2004. 37

[92] J. Riegel, W. Mayer, and Y. van Havre, "Freecad," *Freecadspec2002. pdf*, 2016. 37