Masters Program in Geospatial Technologies

HERI3D: A Comparative Analysis of Traditional and Deep Learning-Based 3D Reconstruction Techniques Using UAV Imagery for Cultural Heritage

Ting-Jia Guo

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS INFORMATION MANAGEMENT SCHOOL
UNIVERSITAT JAUME·I
Universität Münster

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

HERI3D: A Comparative Analysis of Traditional and Deep Learning-Based 3D Reconstruction Techniques Using UAV Imagery for Cultural Heritage

by

Ting-Jia Guo

Master Dissertation presented as partial requirement for obtaining the Master's Degree in Geospatial Technologies

Supervised by

Prof. Benjamin Risse, Institute for Geoinformatics, University of Münster

Constanza Andrea Molina Catricheo, Institute for Geoinformatics, University of Münster

Prof. Sergio Trilles Oliver, Institute of New Imaging Technologies, Universitat Jaume I

February 2026

STATEMENT OF INTEGRITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.
This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.
Münster, 2026
Ting-Jia Guo

USE OF GENERATIVE ARTIFICIAL INTELLIGENCE

Tasks	NO	YES	Generative Artificial Intelligence tools
Better understand issues related to the research		V	SciSpace
Summarizing text from bibliography / resources		V	UniGPT
Summarizing the method(s) used	V		
Translating text		V	Google Gemini
Grammar check		V	Google Gemini
Paraphrase or rewriting text from other people / resources		V	Google Gemini
Coding in R, Python, etc.		V	ChatGPT
Get help on a software		V	ChatGPT
Creating and editing images, maps, videos, etc.	V		
Data analysis	V		
Specify below other tasks not mentioned above:			

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to Prof. Benjamin Risse. I am sincerely grateful for his continuous encouragement and professional support throughout this journey. His suggestions often addressed perspectives I had overlooked, and his attention to these critical details enabled me to continually refine and optimize this research. I also thank Prof. Sergio Trilles Oliver for his support during this program. I would also like to extend my heartfelt thanks to Coni. She is the primary reason I persevered on this research path; her belief in my capabilities gave me the confidence, and I truly did. Whenever I encountered obstacles, her guidance was indispensable in sharpening my research logic and overcoming technical challenges. My appreciation goes to all the members of Prof. Risse's lab. The weekly colloquium was a cornerstone of my growth, helping me develop a consistent research output habit and significantly improving my stability and confidence in academic presentations. I am also grateful to IfGI and the Faculty of Theology. The student assistant opportunities offered the essential financial and structural support that enabled me to complete this thesis. To my dear friend, Hala: thank you for being such an incredible partner on this Master’s journey. Sharing the struggles and cheering each other on made all the difference during the most challenging times. I want to thank my family, who have always provided warm support. Finally, I want to thank myself. My enduring passion for UAV and cultural heritage has been the ultimate driving force behind this work. It was this curiosity about the intersection of technology and history that supported me until the very end.

HERI3D: A Comparative Analysis of Traditional and Deep Learning-Based 3D Reconstruction Techniques Using UAV Imagery for Cultural Heritage

ABSTRACT

This thesis presents a comprehensive comparative analysis of traditional and deep learning-based 3D reconstruction techniques using Unmanned Aerial Vehicle imagery for cultural heritage documentation. The research evaluates four distinct reconstruction paradigms: a traditional Structure-from-Motion and Multi-View Stereo pipeline implemented in COLMAP, neural implicit surface reconstruction using Neuralangelo, radiance field representation via 3D Gaussian Splatting, and a feed-forward geometry-grounded Transformer model known as VGGT. The study uses datasets from three architecturally diverse castle sites in North Rhine-Westphalia, Germany: Schloss Münster, Burg Lüdinghausen, and Schloss Raesfeld. To ensure a fair comparison, a unified evaluation framework was established. This framework incorporates standardized image preprocessing, point cloud refinement, and geometric registration against airborne LiDAR reference data. The performance of each method was assessed through visual qualitative analysis and quantitative evaluation metrics, including Root Mean Square error for accuracy, Cloud-to-Cloud distance for completeness, and local geometric feature descriptors. The results demonstrate that traditional photogrammetry implemented in COLMAP remains the most reliable method for geometric accuracy. Among the learning-based approaches, VGGT with a moderate image count of 24 images consistently achieved the highest completeness and a balanced trade-off between accuracy and geometric stability across all sites. While 3D Gaussian Splatting provides superior visual continuity and color consistency, increasing the number of training iterations primarily refines surface appearance. Neuralangelo maintained global shape continuity but tended to smooth or underrepresent fine-scale architectural details. The findings highlight that reconstruction performance is strongly mediated by site-specific factors, including architectural complexity, UAV flight constraints, and the availability of reference data. This study contributes a reproducible comparative framework that serves as a structured reference for future digital heritage preservation efforts. The results emphasize that no single method dominates across all evaluation criteria and that method selection should align with specific documentation objectives.

KEYWORDS

3D Reconstruction; UAV Photogrammetry; Cultural Heritage; Deep Learning; Comparative Framework

SUSTAINABLE DEVELOPMENT GOALS (SGD):

Table of Contents

INDEX OF FIGURES

INDEX OF TABLES

ACRONYMS

UAV — Unmanned Aerial Vehicle

SfM — Structure from Motion

MVS — Multi-View Stereo

LiDAR — Light Detection and Ranging

ICP — Iterative Closest Point

C2C — Cloud-to-Cloud (distance)

RMS — Root Mean Square (error)

COLMAP — Structure-from-Motion and Multi-View Stereo pipeline

NA — Neuralangelo

3DGS — 3D Gaussian Splatting

VGGT — View-Guided Gaussian Transformer

Comp. — Completeness

1 Introduction

This chapter discusses the technological and practical context of Unmanned Aerial Vehicle (UAV)-based photogrammetry, identifies existing limitations in current reconstruction and evaluation practices, and outlines the research objectives and questions that guide the subsequent analysis.

1.1 Background and Motivation

Three-dimensional reconstruction has become a fundamental approach in cultural heritage preservation, as it enables the precise and long-term documentation of historically and architecturally significant structures. Early studies integrating terrestrial laser scanning and close-range photogrammetry demonstrated that 3D models can capture complex geometry beyond what traditional two-dimensional records can provide (Beraldin, 2004; Pritchard et al., 2017; Owda et al., 2018; Pritchard et al., 2023). Such digital documentation plays a crucial role in safeguarding cultural assets against irreversible loss caused by aging, environmental impacts, or unforeseen damage. International frameworks, such as the UNESCO/PERSIST guidelines (UNESCO/PERSIST, 2016), further emphasize the importance of high-quality digital records as part of long-term heritage preservation strategies. Together, these works establish 3D reconstruction as not merely a visualization tool, but a core component of digital cultural heritage stewardship.

Unmanned Aerial Vehicles (UAVs) have become an increasingly important tool in cultural heritage documentation due to their ability to capture perspectives that are inaccessible from ground-based observation. Aerial viewpoints enable the detailed inspection of roof structures, upper façades, and structural elements that are often difficult or unsafe to access through traditional surveying methods. When combined with photogrammetric techniques such as Structure-from-Motion (SfM) and Multi-View Stereo (MVS), UAV imagery enables flexible and cost-efficient three-dimensional mapping of heritage sites. As a result, UAV-based photogrammetry has been widely adopted for documentation, monitoring, and visualization purposes (Remondino et al., 2011; Nex & Remondino, 2014).

The motivation for this research is grounded in sustained hands-on UAV field experience. Initial UAV practice conducted in Portugal focused primarily on landscape documentation, with limited emphasis on cultural heritage. A decisive shift occurred during the author's Erasmus Mundus studies in Germany, where UAV documentation was deliberately redirected toward cultural heritage sites. A formative UAV mission at Schloss Drachenburg

highlighted that UAV flight is not merely a data acquisition process, but a critical means of integrating UAV technology with three-dimensional reconstruction techniques to support cultural heritage preservation. Motivated by this experience, the author conducted extensive UAV documentation across nearly forty cultural heritage sites in Germany and neighboring regions, including Schloss Glücksburg, Burg Rheinstein, and Château de Vianden. These field activities ultimately formed the basis of this master's thesis research.

Despite recent advances in learning-based reconstruction techniques, systematic and reproducible comparisons between traditional photogrammetric pipelines and emerging methods remain limited, particularly in cultural heritage contexts. This study addresses this gap by evaluating multiple reconstruction paradigms under consistent UAV data acquisition conditions, aiming to inform both methodological development and applied heritage documentation practice.

1.2 Research Objectives and Research Questions

Building upon the practical challenges identified in UAV-based cultural heritage documentation and the emerging diversity of 3D reconstruction paradigms, this study pursues the following objectives:

1. To compare the visual quality of models generated by traditional photogrammetric methods and deep learning–based reconstruction methods.
2. To evaluate and compare different reconstruction methods in terms of accuracy, completeness, and selected geometric features.
3. To establish a reproducible comparative framework that integrates UAV data acquisition, multiple reconstruction pipelines, and standardized evaluation metrics, serving as a reference for future digital heritage preservation.

Guided by these objectives, this research addresses the following questions:

1. What are the differences in visual quality between traditional photogrammetric methods and deep learning–based reconstruction methods?
2. How do different methods perform in terms of model accuracy, completeness, and geometric features?
3. How can a reproducible comparative framework be established to serve as a reference for future digital preservation?

2 Related Work

This chapter reviews existing research on UAV-based 3D reconstruction and its application to cultural heritage documentation, organized into three parts: UAV-based reconstruction for cultural heritage; traditional and learning-based reconstruction paradigms; and evaluation metrics and comparative studies.

2.1 UAV-Based 3D Reconstruction for Cultural Heritage

This section reviews the application of UAV-based 3D reconstruction in cultural heritage documentation, with an emphasis on its role in recording, analyzing, and preserving architectural heritage.

2.1.1 Importance of 3D Reconstruction for Cultural Heritage

3D reconstruction has become a fundamental approach for documenting, preserving, and managing cultural heritage, enabling the creation of accurate and durable digital records of historically significant sites. Cultural heritage assets are increasingly threatened by natural hazards, environmental degradation, urban development, and human-induced destruction, making systematic digital preservation an urgent necessity. High-resolution 3D models preserve geometric and material information even when physical structures are partially damaged or lost.

Compared to traditional 2D plans and sketches, which often lack spatial consistency and semantic richness, 3D models support scalable visualization, precise measurement, and multidisciplinary interpretation, particularly for conservation planning and architectural analysis (Owda et al., 2018). Advances in terrestrial laser scanning (TLS) and photogrammetric techniques have further improved the quality and reliability of heritage documentation. TLS enables the acquisition of dense and accurate point clouds, providing objective as-built records of complex architectural structures. Large-scale documentation projects, such as Cologne Cathedral and Aachen Cathedral, demonstrate how comprehensive 3D datasets support long-term monitoring, conservation decision-making, and structural assessment of UNESCO World Heritage sites (Pritchard et al., 2017; Pritchard et al., 2023). Beyond geometric representation, 3D reconstruction plays a key role in digital knowledge integration.

The integration of 3D models with semantic frameworks such as Historic Building Information Modeling (HBIM) and 3D Geographic Information Systems (3D GIS) enables structured storage and interpretation of historical, architectural, and conservation-related

information (Dore & Murphy, 2012). Hybrid approaches that combine multiple data acquisition techniques have been shown to improve model completeness and accuracy, particularly through the fusion of TLS, photogrammetry, and UAV-based data to address occlusions and accessibility limitations (Zachos & Anagnostopoulos, 2023; Balestrieri et al., 2024).

2.1.2 UAV-Based 3D Reconstruction

Unmanned Aerial Vehicles (UAVs) have become widely adopted platforms for 3D reconstruction due to their flexibility, cost-effectiveness, and ability to acquire high-resolution imagery at low altitudes. Compared to conventional manned aerial photogrammetry, UAV-based image acquisition enables finer-scale geometric documentation, making it particularly suitable for applications requiring detailed spatial representation, such as cultural heritage and architectural documentation (Remondino et al., 2011; Nex & Remondino, 2014).

A typical UAV-based 3D reconstruction workflow follows a photogrammetric pipeline including flight planning, image acquisition with sufficient overlap, camera calibration, image orientation, dense point cloud generation, and textured surface reconstruction. Previous studies demonstrated that UAV imagery acquired with consumer-grade cameras can be processed in a largely automated manner to generate digital surface models, orthophotos, and textured 3D models with acceptable metric accuracy (Remondino et al., 2011). Subsequent reviews further systematized this workflow and confirmed its applicability across geomatics domains, including archaeology, architecture, and cultural heritage (Nex & Remondino, 2014).

Structure-from-Motion (SfM) has become the standard approach for image orientation and sparse reconstruction in UAV-based 3D modeling. SfM simultaneously estimates camera poses and scene geometry from overlapping images without requiring precise prior sensor information, which is particularly advantageous for UAV platforms with limited onboard navigation accuracy (Colomina & Molina, 2014; Jiang et al., 2020). As UAV image datasets continue to grow in size and resolution, conventional SfM pipelines face increasing computational challenges, especially in feature matching and bundle adjustment. The efficiency and scalability of SfM algorithms have therefore become critical limiting factors in large-scale UAV-based reconstruction workflows (Jiang et al., 2020).

2.1.3 UAV-Based 3D Reconstruction for Cultural Heritage

UAV-based photogrammetry for cultural heritage documentation builds upon earlier developments in close-range and architectural photogrammetry, which demonstrated the feasibility of producing detailed and metrically reliable 3D models of historic structures. Image-based reconstruction has been shown to be particularly effective for complex heritage objects, such as castles, where comprehensive geometric documentation is required for visualization, conservation, and historical analysis (Kersten et al., 2024).

The introduction of UAV platforms expanded heritage documentation beyond terrestrial approaches by enabling aerial perspectives, allowing roofs, upper façades, and otherwise inaccessible architectural elements to be captured. Comparative studies indicate that UAV-derived models can achieve geometric accuracies comparable to terrestrial photogrammetry and terrestrial laser scanning (TLS) when appropriate flight planning and control point strategies are applied (Bolognesi et al., 2014).

To improve model completeness and accuracy, hybrid recording strategies integrating UAV imagery with terrestrial photogrammetry and laser scanning have been widely adopted. These approaches exploit the complementary strengths of different sensors, with UAVs covering elevated areas and terrestrial methods providing detailed ground-level information. Case studies on historic castles show that integrated datasets result in more comprehensive 3D models, supporting architectural interpretation and long-term preservation planning (Pattee et al., 2015).

Recent review studies emphasize that UAV platforms combined with SfM–MVS pipelines have become a standard workflow for cultural heritage documentation. While best practices for data acquisition and processing have been established, persistent challenges remain regarding accuracy assessment, scalability, and reproducibility across sites and software environments (Pepe et al., 2022). As a result, accuracy evaluation has emerged as a central research topic. Empirical studies demonstrate that UAV-derived accuracy is strongly influenced by ground control point distribution, camera calibration, and georeferencing strategy, with centimeter-level accuracy achievable under favorable conditions, highlighting the need for rigorous validation when UAV photogrammetry is used for scientific documentation and conservation decision-making (Günen et al., 2024).

2.2 Traditional and Deep Learning-Based 3D Reconstruction

This section reviews traditional geometric and learning-based 3D reconstruction approaches, covering classical Structure-from-Motion and Multi-View Stereo pipelines as well as recent neural scene representations and geometry-grounded learning frameworks.

2.2.1 Traditional SfM/MVS-based Methods

Traditional Structure-from-Motion (SfM) and Multi-View Stereo (MVS) methods constitute the foundational paradigm for image-based 3D reconstruction. The mathematical formulation of camera models, epipolar geometry, triangulation, and projective reconstruction provides the core principles that enable the recovery of camera poses and sparse scene structure from image correspondences (Hartley & Zisserman, 2004).

Among different SfM strategies, incremental SfM has become the most widely adopted approach due to its robustness and high reconstruction accuracy. Incremental methods initialize the reconstruction from a carefully selected image pair and progressively register additional images while continuously refining camera poses and 3D structure via bundle adjustment. Despite their strong performance, incremental pipelines are computationally demanding and may suffer from error accumulation, sensitivity to initialization, and scalability limitations when applied to very large image datasets (Schönberger & Frahm, 2016).

In contrast to incremental pipelines, global SfM approaches have been proposed as an alternative paradigm. Recent advances demonstrate that modern global SfM systems can achieve accuracy and robustness comparable to state-of-the-art incremental pipelines while significantly improving computational efficiency. These developments indicate a convergence between incremental and global paradigms, while also exposing persistent challenges related to noise sensitivity, degenerate configurations, and reliance on handcrafted feature matching (Pan et al., 2024).

Following the estimation of camera poses and sparse scene structure through SfM, dense Multi-View Stereo (MVS) serves as the key mechanism for recovering explicit geometric detail by establishing dense correspondences across multiple calibrated images. Early advances in dense matching efficiency were enabled by randomized correspondence algorithms, most notably PatchMatch, which introduced an iterative propagation and random search strategy to approximate dense pixel-level correspondences with significantly reduced computational complexity compared to exhaustive search (Barnes et al., 2009). Building upon this concept, later multiview extensions reformulated PatchMatch-based matching in

scene space and incorporated surface normal estimation to enable robust multi-view depth inference. Massively parallel MVS approaches demonstrated that dense depth maps and surface normals could be efficiently recovered by aggregating photo-consistency across multiple views, while maintaining high accuracy and scalability through GPU-based parallelization (Galliani et al., 2015). These methods established dense MVS as a standard component of traditional SfM/MVS pipelines, particularly for applications requiring high-resolution surface reconstruction.

2.2.2 Learning-Based Depth and Multi-View Stereo

Traditional multi-view stereo methods rely on handcrafted similarity metrics and explicit regularization schemes, which often struggle under challenging conditions such as low-textured surfaces, specular reflections, and complex illumination. Motivated by advances in deep learning for stereo matching, recent research has increasingly explored learning-based approaches to replace or augment conventional MVS pipelines by learning robust feature representations and matching costs directly from data (Stathopoulou & Remondino, 2023). Recent survey studies categorize learning-based MVS methods according to scene representation, cost volume construction strategies, and the integration of geometric priors. Among these, depth-map-based approaches, exemplified by MVSNet and its successors, are considered particularly scalable for large-scale photogrammetric applications due to reduced memory requirements and compatibility with classical SfM pipelines (Stathopoulou & Remondino, 2023). Nevertheless, learning-based MVS methods remain challenged by generalization across datasets, sensitivity to training data bias, and computational demands during inference, motivating continued exploration of hybrid geometric–learning approaches.

Beyond depth estimation, deep learning has also been introduced to enhance feature correspondence estimation, a critical component of both traditional SfM and learning-based MVS pipelines. SuperGlue formulates feature matching as a learnable optimization problem using graph neural networks and differentiable optimal transport, enabling robust correspondence assignment and outlier rejection under challenging conditions such as large viewpoint changes and low-texture regions (Sarlin et al., 2020). More recent work extends learning-based matching to dense correspondence estimation. RoMa combines robust foundation-model features with fine-grained representations to achieve dense matching across extreme appearance variations, highlighting the growing role of dense

correspondence learning as a bridge between classical SfM and modern neural reconstruction pipelines (Edstedt et al., 2024).

2.2.3 Neural Scene Representation

Recent advances in deep learning have introduced neural scene representations that differ fundamentally from traditional explicit geometric models. These approaches encode geometry and appearance within learnable representations, enabling high-quality novel view synthesis and photorealistic rendering directly from image observations.

2.2.3.1 Neural Radiance Fields

Neural Radiance Fields (NeRF) introduced a seminal neural scene representation that models a static scene as a continuous volumetric function parameterized by a multilayer perceptron. Given a 5D input of 3D spatial coordinates and viewing direction, the network predicts volume density and view-dependent radiance, which are integrated along camera rays using differentiable volume rendering to synthesize novel views (Mildenhall et al., 2020). By encoding geometry and appearance implicitly, NeRF enables highly photorealistic novel view synthesis while avoiding the memory limitations of discretized voxel representations. However, NeRF assumes static geometry and consistent illumination, limiting its applicability to unconstrained real-world imagery. NeRF in the Wild (NeRF-W) addresses these limitations by introducing latent appearance embeddings and separating static and transient scene components, allowing the model to handle illumination variation, transient objects, and exposure changes. This extension enables neural scene reconstruction from large, unstructured photo collections of cultural landmarks and significantly broadens the applicability of neural radiance field methods to real-world scenarios (Martin-Brualla et al., 2021).

2.2.3.2 Frameworks and Systems for Neural Scene Representation

As NeRF-based methods rapidly evolved, the absence of unified tooling and reproducible pipelines became a practical challenge. Nerfstudio addresses this gap by providing a modular framework that integrates data preprocessing, model development, rendering, and interactive visualization within a unified system. By supporting multiple NeRF variants and standardized export formats, Nerfstudio has accelerated research, improved reproducibility, and facilitated the application of neural scene representations to real-world datasets (Tancik et al., 2023).

2.2.3.3 Gaussian-Based Scene Representation

Despite their strong representational capacity, NeRF-based methods remain computationally expensive due to volumetric ray marching and implicit neural inference. To address these limitations, recent work has introduced explicit neural scene representations based on 3D Gaussian primitives. 3D Gaussian Splatting represents scenes as collections of anisotropic 3D Gaussians optimized directly from images and rendered through a fast, differentiable splatting process (Kerbl et al., 2023). Building on this representation, SuGaR incorporates surface-aligned regularization to encourage Gaussians to follow underlying scene geometry, enabling efficient extraction of high-quality surface meshes via Poisson reconstruction (Guédon & Lepetit, 2023).

Recent survey studies provide comprehensive overviews of the rapidly expanding literature on 3D Gaussian Splatting, highlighting advances in optimization, scalability, and compression. These works position Gaussian-based representations as a promising compromise between implicit neural fields and explicit geometric models, offering favorable trade-offs between rendering quality, computational efficiency, and scene editability (Dalal et al., 2024; Luo et al., 2024; Ali et al., 2025).

2.2.4 Geometry-Grounded Learning

Recent advances in 3D reconstruction have led to geometry-grounded learning approaches that explicitly integrate geometric principles into learning-based frameworks. A representative example is Neuralangelo, which combines neural implicit surface representations with classical multi-view geometry constraints to achieve high-fidelity surface reconstruction (Li et al., 2023).

Beyond surface reconstruction, geometry-grounded learning has been extended to the Structure-from-Motion pipeline itself. VGGSfM proposes a fully differentiable and geometry-aware SfM framework that unifies feature tracking, camera estimation, triangulation, and bundle adjustment within an end-to-end architecture. By replacing discrete solvers with learnable modules and differentiable bundle adjustment, VGGSfM retains the core geometric structure of classical SfM while enabling joint optimization of all components, achieving strong performance on challenging benchmarks (Wang et al., 2024).

More recent feed-forward and large-scale models further generalize this paradigm. VGGT introduces a transformer-based architecture that jointly predicts camera parameters, depth maps, point maps, and point tracks in a single forward pass, treating geometry as a learned latent structure while preserving geometric consistency (Wang et al., 2025). In parallel,

DUSt3R regresses dense point maps directly from image pairs without requiring known camera parameters, implicitly encoding geometry and camera relationships while enabling multi-view consistency through global alignment (Wang et al., 2024). Building on this idea, MapAnything extends geometry-grounded learning to heterogeneous inputs, demonstrating a unified framework capable of bridging classical SfM, MVS, and modern neural inference (Keetha et al., 2025).

2.3 Evaluation Metrics and Comparison Studies

This section reviews commonly adopted evaluation metrics and comparative frameworks for assessing 3D reconstruction quality. Quantitative and qualitative evaluation criteria are considered, beginning with geometric accuracy and distance-based metrics, followed by comparative studies across different reconstruction technologies.

2.3.1 Geometric Accuracy and Distance-Based Metrics

Geometric accuracy is a fundamental criterion for evaluating 3D reconstruction quality, particularly in cultural heritage documentation, conservation, and analysis, where even small geometric deviations may compromise metric interpretation and long-term preservation (Beraldin, 2004).

A seminal comparative study assessed multi-view stereo algorithms using point-to-surface and surface-to-surface distances against high-accuracy reference models, jointly considering reconstruction accuracy and completeness to enable balanced comparison across methods and datasets (Seitz et al., 2006). However, simple closest-point distance measures are sensitive to surface roughness, point density variations, and registration errors. To address these limitations, subsequent work introduced normal-based distance computation and uncertainty-aware metrics that explicitly incorporate surface roughness and confidence intervals, improving robustness in complex and irregular scenes (Lague et al., 2013).

In cultural heritage applications, geometric accuracy must often be evaluated alongside semantic and representational requirements. A review of heritage Building Information Modeling (HBIM) highlighted that while high geometric fidelity is essential for documentation and visualization, geometry-only metrics may not fully capture the semantic richness required for conservation and management workflows. This highlights an inherent trade-off between geometric accuracy and semantic abstraction, underscoring the need for application-aware evaluation strategies (Radanovic et al., 2020).

2.3.2 Comparative Studies in Cultural Heritage Reconstruction

Recent years have seen an increasing number of comparative studies evaluating different 3D reconstruction paradigms for cultural and architectural heritage.

A representative study by Clini et al. compares Structure-from-Motion with Multi-View Stereo (SfM–MVS), Neural Radiance Fields (NeRF), and 3D Gaussian Splatting (GS) for architectural heritage documentation using low-cost sensors (Clini et al., 2024). Their results reveal a clear trade-off between geometric accuracy and visual expressiveness. While NeRF and GS outperform SfM–MVS in rendering quality and processing efficiency, SfM–MVS consistently achieves higher geometric accuracy when evaluated against terrestrial laser scanning (TLS) reference data. In particular, NeRF exhibits increased noise on planar or weakly textured surfaces, and GS shows limited control over geometric fidelity, indicating that neural rendering approaches remain insufficient as standalone solutions for metric heritage documentation.

Complementary insights are provided by sensor-based comparisons. Gaong et al. evaluate UAV-based photogrammetry against TLS for building reconstruction, demonstrating TLS as a benchmark for geometric accuracy due to its dense and uniform point sampling (Gaong et al., 2025). While UAV photogrammetry typically achieves centimeter-level accuracy under appropriate conditions, TLS delivers millimeter- to sub-centimeter precision, reinforcing its role as a reference standard for assessing image-based and learning-based reconstruction methods.

Beyond direct comparison, recent work explores integrative evaluation frameworks. Yu et al. propose a workflow-oriented comparison between 3D Gaussian Splatting and Light Detection and Ranging (LiDAR)-based point clouds for modern architectural heritage, emphasizing functional complementarity rather than methodological competition (Yu et al., 2025). Their findings suggest positioning GS as a visualization and interaction layer built upon geometrically reliable LiDAR data, rather than as a replacement for established surveying workflows.

3 Dataset and Study Area

The chapter introduces the characteristics of the study areas, details the UAV imagery data acquisition and flight planning strategy, and describes the airborne LiDAR reference data used for quantitative and comparative evaluation.

3.1 Cultural Heritage Sites

To evaluate the performance of UAV-based photogrammetric 3D reconstruction techniques for cultural heritage documentation, three castle sites located in North Rhine-Westphalia (NRW), Germany—Münster, Lüdinghausen, and Raesfeld—were selected and recorded using UAV imagery and reconstructed in 3D within the scope of this study.

The baroque palace in Münster (Fig. 1, left), located in the city centre of Münster, North Rhine-Westphalia, was selected as one of the study sites. Constructed in the late eighteenth century as the residence of the prince-bishop, the palace is characterized by its large-scale, symmetrical façade and ordered architectural layout. These regular geometric features make the site suitable for assessing reconstruction performance.

Burg Lüdinghausen (Fig. 1, center) is a medieval moated castle located in the town of Lüdinghausen, North Rhine-Westphalia. Constructed in the thirteenth century, the castle features a compact layout with enclosed courtyards, and surrounding vegetation. These characteristics create increased geometric complexity and occlusion effects, providing a contrasting test case for evaluating reconstruction completeness and structural detail preservation.

Schloss Raesfeld (Fig. 1, right), located in the municipality of Raesfeld, North Rhine-Westphalia, is a historically significant Renaissance moated castle. Originally constructed in the early fourteenth century and extensively remodeled in the sixteenth century, the castle combines defensive fortification elements with Renaissance residential architecture. The presence of surrounding water bodies, vertical walls, and complex roof structures introduces challenges related to reflections, occlusions, and viewpoint limitations, making the site well suited for evaluating reconstruction robustness in more complex environmental settings.

Together, the three selected sites represent a diverse range of architectural styles, and environmental conditions, enabling a comprehensive evaluation of UAV-based 3D reconstruction methods across varying cultural heritage contexts.

Fig. 1 Schloss Münster (left), Burg Lüdinghausen (center), and Schloss Raesfeld (right).

3.2 UAV Data Acquisition and Flight Planning

UAV imagery data acquisition was conducted with the objective of ensuring sufficient geometric coverage and viewpoint diversity to support reliable Structure-from-Motion (SfM) reconstruction and subsequent comparative evaluation of traditional and deep learning-based 3D reconstruction methods. Attention was given to the distribution of camera viewpoints, image overlap, and camera tilt angles, as these factors are known to critically influence reconstruction robustness and completeness.

All reconstruction methods evaluated in this study were applied to the same UAV imagery datasets acquired under this flight configuration to ensure a controlled and fair comparison across reconstruction paradigms. All UAV imagery was acquired using a DJI Mini 3 platform, which was selected as the experimental UAV system for all flight missions conducted in this study, as illustrated in Fig. 2.

Fig. 2 DJI Mini 3 (left). Drone planned route (right)

Following established best practices in UAV-based photogrammetry, the imagery acquisition strategy combined circular flight trajectories with oblique camera orientations to capture both roof structures and vertical façades. Previous studies have demonstrated that circular flight paths with varying camera tilt angles significantly improve feature visibility and reduce occlusion effects, especially for complex architectural forms (Jo et al., 2025). In addition, multi-elevation and oblique viewing configurations enhance viewpoint continuity, which is essential for stable camera registration and dense 3D reconstruction.

In this study, UAV flights were planned to surround each castle site at a roughly constant radius, while systematically varying the camera tilt angle relative to the horizontal plane. As illustrated in Fig. 3, images were captured using oblique viewing angles of approximately −20°, −30°, and −40°, allowing the camera to simultaneously observe roof surfaces, upper façades, and vertical wall structures, thereby supporting feature continuity across elevations while minimizing extreme perspective distortion.

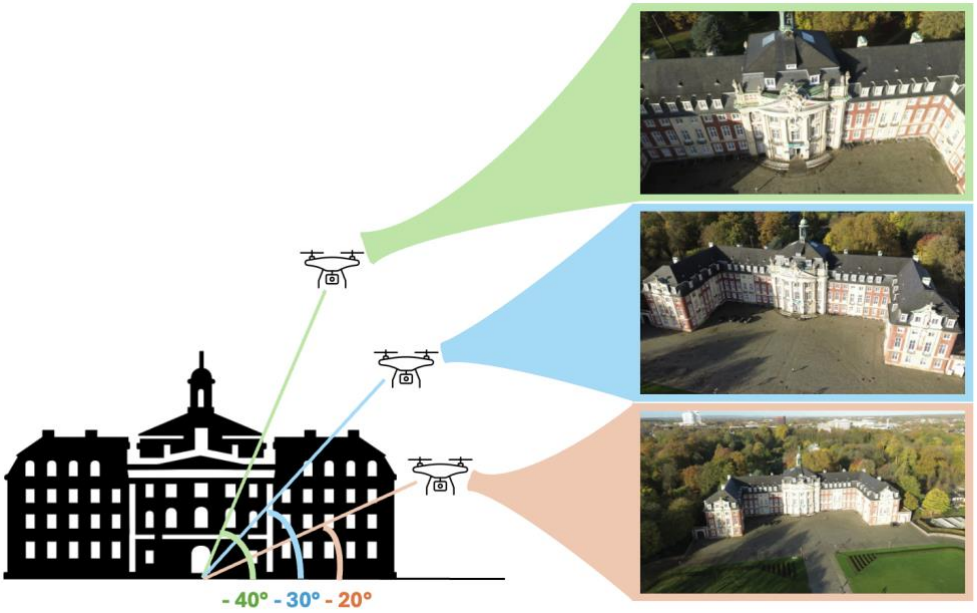

Fig. 3 UAV imagery data acquisition strategy.

The adopted acquisition design is consistent with findings from recent quantitative analyses of UAV flight parameters for SfM-based building reconstruction, which highlight the importance of oblique imaging and sufficient overlap for improving reconstruction completeness and reducing reprojection error (Jo et al., 2025). Moreover, ensuring continuous viewpoint transitions across elevations is particularly relevant for downstream learning-based reconstruction methods, such as Neural Radiance Fields and 3D Gaussian Splatting (Ham et al., 2024).

All imagery datasets were collected under stable illumination conditions to reduce radiometric inconsistencies. The resulting UAV imagery provides a consistent and reproducible dataset for evaluating performance differences between traditional photogrammetric pipelines and deep learning-based 3D reconstruction methods.

3.3 LiDAR Reference Data

To support the geometric evaluation of the reconstructed 3D models, official airborne LiDAR data provided by the state of North Rhine-Westphalia (NRW), Germany, were used

as reference data in this study. The LiDAR datasets were obtained from the Geobasis NRW open geospatial data portal and represent state-wide, high-resolution elevation measurements acquired through airborne laser scanning.

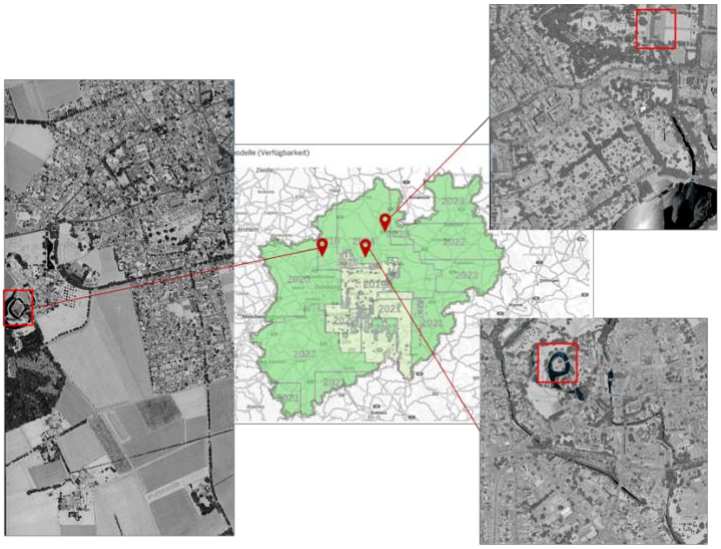

Fig. 4 Spatial distribution of heritage sites, airborne LiDAR reference in NRW.

As illustrated in Fig. 4, the selected castle sites are located within the spatial coverage of the LiDAR datasets, enabling direct geometric correspondence between the UAV-based reconstructions and the reference data. The LiDAR point clouds provide dense and consistent geometric information at the landscape and building scales, making them suitable for evaluating overall geometric accuracy and structural completeness.

The public LiDAR data were delivered with standard point cloud classification, from which building-related classes were extracted as the initial reference geometry. To ensure correspondence with the cultural heritage sites, the extracted building point clouds were further manually refined to remove surrounding non-heritage structures and unrelated built elements. This refinement step ensured that the LiDAR reference data primarily represented the target cultural heritage structures.

It should be noted that the LiDAR data are not treated as absolute ground truth at the level of fine architectural detail, but rather as a reliable geometric reference for large-scale structural comparison. Differences in acquisition geometry, point density, and sensing modality between airborne LiDAR and image-based reconstruction methods are therefore accounted for.

4 Methodology

This study proposes a unified comparative evaluation framework to assess UAV-based 3D reconstruction methods across heterogeneous architectural heritage sites. The framework integrates four reconstruction paradigms: a traditional SfM/MVS baseline (COLMAP), neural surface reconstruction (Neuralangelo), radiance field representation (3D Gaussian Splatting), and feed-forward geometry-grounded Transformers (VGGT). Each method is treated as an independent reconstruction pipeline operating on site-specific datasets, followed by standardized post-processing steps. As illustrated in Fig. 5, all reconstruction outputs undergo a unified workflow.

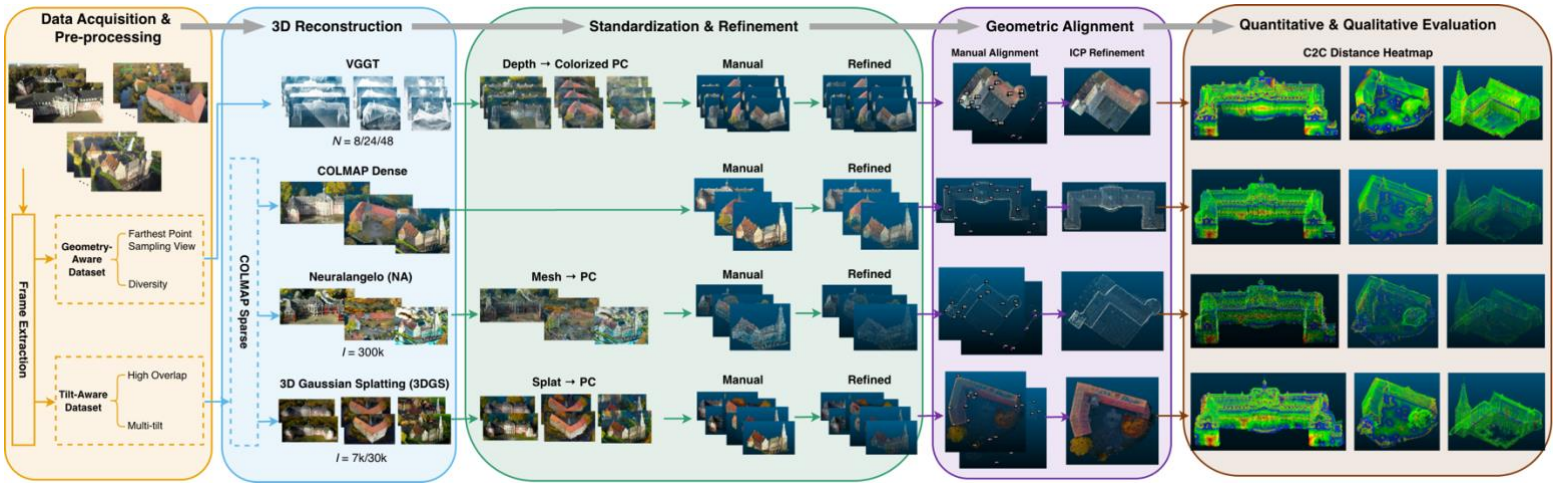

Fig. 5 Overview of the HERI3D comparative evaluation framework. This study integrates UAV data acquisition, four 3D reconstruction paradigms, and a standardized geometric evaluation pipeline.

The evaluation is performed independently for each heritage site (M, L, and R), allowing site-specific characteristics. By enforcing consistent preprocessing, cleaning, and alignment strategies prior to evaluation, the proposed framework isolates reconstruction behavior from methodological bias.

4.1 Image Pre-processing

Following the data acquisition phase, the raw video sequences recorded by the Unmanned Aerial Vehicle (UAV) must be converted into high-quality image frames suitable for photogrammetric and neural reconstruction. This pre-processing stage is critical, as the density and quality of the extracted frames directly influence the robustness of the Structure-from-Motion (SfM) algorithms and the convergence of neural reconstruction models.

4.1.1 Tilt-Aware Frame Extraction under Different UAV Flight Strategies

In this study, two distinct UAV flight strategies were employed for cultural heritage documentation: orbital circular missions and feature-specific missions**.** Owing to differences in camera pose variability and gimbal behavior, a unified frame extraction strategy is insufficient to ensure consistent spatial coverage and geometric reliability across all missions. The spatial ground coverage of a video frame is jointly determined by UAV altitude, camera field of view, and gimbal tilt angle. Under a simplified pinhole camera assumption, the effective ground footprint along the viewing direction can be approximated as a function of the UAV altitude H, the camera vertical field of view VFOV, and the gimbal tilt angle θ. The near and far ground limits of the footprint can be expressed as:

$$D_{\text{near}} = H \cdot \tan\left(\theta - \frac{\text{VFOV}}{2}\right), \quad D_{\text{far}} = H \cdot \tan\left(\theta + \frac{\text{VFOV}}{2}\right)$$

As illustrated in Fig. 6, variations in gimbal tilt lead to non-linear changes in ground footprint extent and frame-to-frame overlap. Consequently, relying on a fixed extraction frequency or a uniform overlap threshold across all flight types may result in either excessive redundancy or the loss of critical viewpoints, depending on the mission geometry. To address this issue, a custom Python pipeline for frame extraction was developed to synchronize UAV flight logs with video timestamps, enabling tilt-aware image extraction.

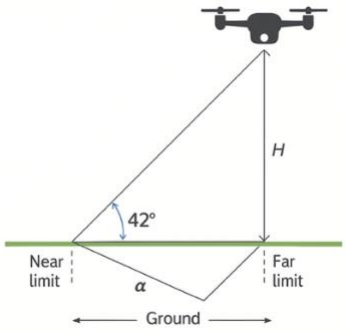

Fig. 6 Geometry of ground footprint variation under different gimbal tilt angles.

1. Orbital Circular Missions — Overlap-Based Extraction.
   For orbital missions, the UAV maintained a constant flight altitude of approximately 40 m while systematically varying the gimbal pitch angle (−20°, −30°, and −40°). Under these controlled conditions, frame extraction was governed by a strict forward-overlap criterion. The expected ground footprint was estimated

using the formulation above, and frames were extracted only when the forward overlap exceeded 90%. This high-redundancy configuration ensures robust feature correspondences and stable camera registration (Schönberger & Frahm, 2016).

2. Feature-Specific Missions — Spacing-Based Extraction.
   In contrast, feature-specific missions focused on localized architectural elements such as façades, towers, and roof details. These maneuvers involved variable flight speeds, abrupt changes in viewing direction, and irregular gimbal motion, resulting in highly inconsistent overlap between consecutive frames. Applying a uniform overlap constraint in such scenarios often led to substantial data loss. To mitigate this issue, a spacing-based extraction strategy was adopted, enforcing a fixed spatial interval of approximately 3 m between consecutive frames. This approach ensures consistent spatial sampling of architectural details (Li et al., 2024).

By separating frame extraction strategies according to flight behavior and explicitly accounting for gimbal-induced footprint variation, the proposed preprocessing pipeline preserves global structural continuity and local geometric detail. This distinction is particularly important for downstream learning-based reconstruction methods, which are sensitive to viewpoint distribution and input sparsity.

4.1.2 Geometry-Aware Dataset Stratification for Transformer-Based Reconstruction

For Visual Geometry Grounded Transformer (VGGT) reconstruction, a dedicated dataset stratification strategy was adopted to accommodate the fundamentally different assumptions of transformer-based models compared to traditional photogrammetric pipelines. Unlike SfM-based methods, which benefit from dense, sequential image overlap, VGGT relies on globally distributed multi-view geometry to establish long-range spatial correspondence (Wang et al., 2025). Consequently, simply increasing the number of input frames does not necessarily improve reconstruction quality and may even degrade performance. To address this, a Farthest Point Sampling (FPS) image selection strategy was implemented. FPS is designed to maximize spatial diversity among selected viewpoints by iteratively choosing images whose camera centers are maximally distant from previously selected ones. The stratification process consisted of two stages:

1. Camera Center Extraction
   A preliminary sparse reconstruction was first generated using COLMAP with the SIMPLE_PINHOLE camera model. From this reconstruction, camera centers

$$\mathbf{c}_i = (x_i, y_i, z_i)$$

were extracted for all candidate frames, providing a metric spatial representation of camera poses.

2. FPS Subset Selection

   Based on the extracted camera centers, FPS was applied to select subsets of increasing size. Starting from an initial randomly selected camera center $\mathbf{c}_0$, FPS iteratively selects the next camera center according to:

$$\mathbf{c}_k = \arg\max_{\mathbf{c}_i} \min_{\mathbf{c}_j \in S_{k-1}} \|\mathbf{c}_i - \mathbf{c}_j\|_2,$$

   where S_{k-1}denotes the set of previously selected camera centers.

   Using this procedure, three subsets consisting of 8, 24, and 48 images were generated.

Preliminary experiments were conducted using substantially larger image sets, with the number of frames increased up to 144. However, empirical results showed that reconstruction quality deteriorated as the number of input images increased beyond a moderate threshold. Accordingly, the 8-image subset represents an extreme sparsity condition to evaluate the lower bound of VGGT performance, the 24-image subset reflects a balanced configuration aligned with prior work, and the 48-image subset serves as an upper-bound test to examine whether additional views yield tangible benefits.

4.2 Traditional Photogrammetry Pipeline

Traditional photogrammetry is employed in this study as a geometry-grounded baseline for 3D reconstruction. The following subsection details the implementation and parameterization of the COLMAP pipeline used in this study.

4.2.1 COLMAP

COLMAP is employed in this study as the representative traditional photogrammetric pipeline for three-dimensional reconstruction. It provides an integrated framework for Structure-from-Motion (SfM) and Multi-View Stereo (MVS), enabling the recovery of camera poses and scene geometry from unordered image collections (Schönberger & Frahm, 2016). To accommodate the computational demands of high-resolution UAV imagery, the reconstruction workflow was executed using the PALMA High-Performance Computing (HPC) infrastructure.

1. Sparse Reconstruction and Camera Pose Estimation

The SfM stage aims to jointly estimate camera parameters and sparse scene geometry. Local image features were extracted using the Scale-Invariant Feature Transform (SIFT), followed by exhaustive pairwise matching to establish feature correspondences. Given the quadratic growth of potential image pairs in large datasets, feature extraction and matching were accelerated using GPU-based computation on the HPC cluster, substantially reducing processing time.

Sparse reconstruction was performed incrementally, allowing camera poses to be registered progressively while maintaining geometric consistency. Camera pose estimation and scene structure were refined through Bundle Adjustment (BA), which minimizes the overall reprojection error across all observed image points. Formally, BA solves the following non-linear least squares problem:

$$\min_{\{\mathbf{R}_i,\mathbf{t}_i,\mathbf{X}_j\}} \sum_{i,j} \left\| \pi(\mathbf{R}_i\mathbf{X}_j + \mathbf{t}_i) - \mathbf{x}_{ij} \right\|^2,$$

where $\mathbf{R}_i$and $\mathbf{t}_i$denote the rotation and translation of camera i, $\mathbf{X}_j$represents a 3D point in the scene, $\mathbf{x}_{ij}$is its observed 2D projection in image i, and $\pi(\cdot)$denotes the camera projection function.

To ensure stable reconstruction under varying viewpoint distributions, conservative geometric constraints were applied during mapping, including a minimum triangulation angle of 4° and a minimum number of inlier correspondences for absolute pose estimation.

2. Headless Execution and Dense Surface Generation in HPC Environments

   A significant technical challenge in deploying COLMAP within high-performance computing (HPC) environments arises from its inherent dependency on graphical libraries such as Qt and OpenGL. On headless compute nodes lacking a physical display or X-server, standard COLMAP distributions frequently encounter runtime exceptions, which can disrupt large-scale or automated reconstruction workflows.

   To overcome these constraints, a specialized development version of COLMAP (v3.14.0.dev0) was compiled from source using the Ninja build system within a dedicated Conda environment. The compilation was explicitly configured with `GUI_ENABLED=OFF` and `OPENGL_ENABLED=OFF`, removing dependencies on graphical interfaces. This design choice allowed the COLMAP binaries to execute natively on PALMA's headless compute nodes without requiring virtual framebuffers (e.g., Xvfb) or additional environment configurations such as `QT_QPA_PLATFORM`. The

resulting custom build, linked against CGAL and SuiteSparse, ensured stable execution of both sparse and dense reconstruction stages while fully utilizing the cluster's multi-threading capabilities.

Following the completion of the sparse reconstruction, the pipeline proceeded to the Multi-View Stereo (MVS) stage for dense point cloud generation. Input images were first processed using `image_undistorter` to normalize lens distortions based on the estimated camera parameters. Depth map estimation was then performed using the Patch-Match Stereo algorithm.

To balance the trade-off between geometric detail preservation and noise suppression, the stereo fusion process was configured with varying levels of aggressiveness depending on site-specific characteristics. For primary datasets such as Burg Lüdinghausen and Schloss Raesfeld, fusion parameters were adjusted to prioritize surface completeness. By modulating parameters such as `StereoFusion.min_num_pixels` and `max_traversal_depth`, the pipeline could generate either a balanced output suitable for general visualization or a more aggressive reconstruction to capture fine architectural details in shadowed or partially occluded regions.

This adaptive parameterization reflects the methodological stance adopted throughout the study: reconstruction quality is not determined by algorithm choice, but by the interaction between scene geometry, data acquisition conditions, and processing configuration. Ensuring stable execution in a headless HPC environment and tailoring dense fusion parameters accordingly were essential steps in producing point clouds with sufficient density and geometric integrity.

4.3 Deep Learning-Based Reconstruction Methods

In this study, three representative learning-based reconstruction methods—Neuralangelo, 3D Gaussian Splatting, and VGGT—are employed to examine how different neural paradigms respond to UAV-based imagery data. The following subsections describe the implementation and methodological characteristics of each approach.

4.3.1 Neuralangelo

Neuralangelo is employed in this study as the representative neural implicit surface reconstruction pipeline. It integrates neural volume rendering with signed distance field (SDF)–based surface optimization to recover dense surface geometry from multi-view

images without auxiliary depth supervision (Li et al., 2023). All Neuralangelo experiments rely on the official default configuration unless explicitly stated otherwise. The complete Neuralangelo workflow was executed and documented as a traceable pipeline on the PALMA High-Performance Computing (HPC) infrastructure.

1. Camera Pre-processing, Undistortion, and Dataset Normalization

   Neuralangelo requires accurate, distortion-free camera intrinsics and consistent per-image extrinsics to ensure stable differentiable rendering and geometric convergence. To satisfy these requirements and to maintain comparability across reconstruction methods, camera parameters were obtained through a COLMAP-based pre-processing stage shared with the traditional photogrammetric pipeline described in Section 4.2.1. All scenes were subsequently normalized using a PINHOLE camera model.

   1) COLMAP sparse reconstruction and image undistortion

      For each dataset, images were first processed in COLMAP to estimate camera poses and intrinsic parameters via sparse reconstruction. The resulting sparse model was then exported through image undistortion, removing radial and tangential lens distortions and producing distortion-free images together with normalized camera intrinsics.

   2) Generation of *transforms.json*

      The undistorted camera parameters were converted into Neuralangelo's JSON-based input representation using the provided `convert_data_to_json.py` script, producing a `transforms.json` file under each dataset directory. This file serves as the single source of truth for camera geometry and scene normalization throughout Neuralangelo training. `transforms.json` encodes the following components:

      - PINHOLE camera intrinsics and image resolution, including focal lengths (f_x, f_y), principal point coordinates (c_x, c_y), and image dimensions (w, h) .
        Example (Schloss Raesfeld): $w = 3759$, $h = 2114$, $f_x = f_y = 2825.603$, $c_x = 1879.5$, $c_y = 1057.0$.
      - Zero distortion coefficients after undistortion, where radial and tangential distortion terms are explicitly set to zero ($k_1 - k_4 =$

$0, p_1 = p_2 = 0)$, confirming that camera geometry is treated as distortion-free during training.

- Per-image extrinsic parameters, represented as a list of frames, each containing an image path and a 4×4 camera-to-world transformation matrix $\mathbf{T}_{c \to w} \in \mathbb{R}^{4 \times 4}$. This explicit encoding ensures deterministic pose alignment across runs.
- Scene bounds and scaling parameters, including axis-aligned bounding boxes and bounding spheres (e.g., `aabb_scale`, `aabb_range`, `sphere_center`, `sphere_radius`), which are used internally to support unbounded outdoor scenes and to enable stable spatial sampling during optimization.

By explicitly recording camera geometry and scene normalization in `transforms.json`, Neuralangelo is constrained to operate on the same camera solution as COLMAP.

2. Implicit Surface Representation and Optimization Formulation

In contrast to photogrammetric pipelines that reconstruct geometry explicitly through triangulation and depth fusion, Neuralangelo represents scene geometry implicitly using a signed distance function (SDF)

$$f : \mathbb{R}^3 \to \mathbb{R},$$

where the reconstructed surface $\mathcal{S}$ is defined as the zero level set

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 \mid f(\mathbf{x}) = 0\}.$$

To enforce the signed distance property and geometric consistency, Neuralangelo introduces an Eikonal regularization term:

$$\mathcal{L}_{\text{eik}} = \mathbb{E}_{\mathbf{x}} \left(\|\nabla f(\mathbf{x})\|_2 - 1 \right)^2 .$$

Following Li et al. (2023), numerical gradients are used to evaluate $\nabla f(\mathbf{x})$, mitigating locality artifacts associated with analytical differentiation over multi-resolution hash encodings and enabling a stable coarse-to-fine optimization process.

The data fidelity term is defined through differentiable volumetric rendering, minimizing the discrepancy between rendered pixel colors $\hat{\mathbf{C}}$ and observed image colors $\mathbf{C}$:

$$\mathcal{L}_{\text{render}} = \mathbb{E}_{\mathbf{r}} \left\| \hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|_2^2,$$

where **r** denotes a camera ray sampled from the input views.

The overall Neuralangelo optimization problem is formulated as a weighted energy minimization:

$$\min_{\theta} \ \mathcal{L}_{\text{render}} + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}} + \lambda_{\text{curv}} \mathcal{L}_{\text{curv}},$$

where θ denotes the network parameters and λ_{eik}, λ_{curv} are fixed regularization weights. The values of these weights were fixed across all scenes and explicitly recorded in the exported `config.yaml`, ensuring consistent geometric regularization throughout the experiments.

3. Configuration Management and Headless Execution in HPC Environments

   To accommodate the computational demands of large-scale outdoor scenes, Neuralangelo training was executed on the PALMA High-Performance Computing (HPC) infrastructure using NVIDIA A100 GPUs. All runs were performed on headless compute nodes without a display server.

   To ensure stable execution in this environment, training was conducted under offscreen-compatible settings, avoiding runtime failures caused by implicit graphical dependencies, analogous to the headless execution strategy adopted for COLMAP.

   Neuralangelo training was governed by two configuration layers:

   1) a per-scene source configuration (`projects/neuralangelo/configs/custom/<scene>.yaml`) defining dataset bindings and core settings, and
   2) an automatically exported effective run-time configuration (`config.yaml`) stored within each run directory.

   The exported `config.yaml` records all training-critical parameters, including optimization schedules, loss weights, ray sampling strategies, and dataset bindings. For each site, the effective configuration specifies a training schedule of 300,000 iterations, checkpointing every 20,000 iterations, AdamW optimization with a learning rate of 1×10^{-3}, and training from scratch without pretrained weights.

4. Checkpoint-based Validation, Model Selection, and Output Generation

   Unlike traditional photogrammetric pipelines, Neuralangelo does not provide an explicit reprojection-error objective comparable to bundle adjustment. Therefore,

validation was integrated directly into the pipeline through checkpoint-based monitoring and geometry-based sanity checks.

1) Checkpoint-based monitoring

   Meshes were extracted at 100k, 200k, and 300k iterations to verify stable convergence and inspect the qualitative evolution of recovered surface geometry, as illustrated in Fig. 7.

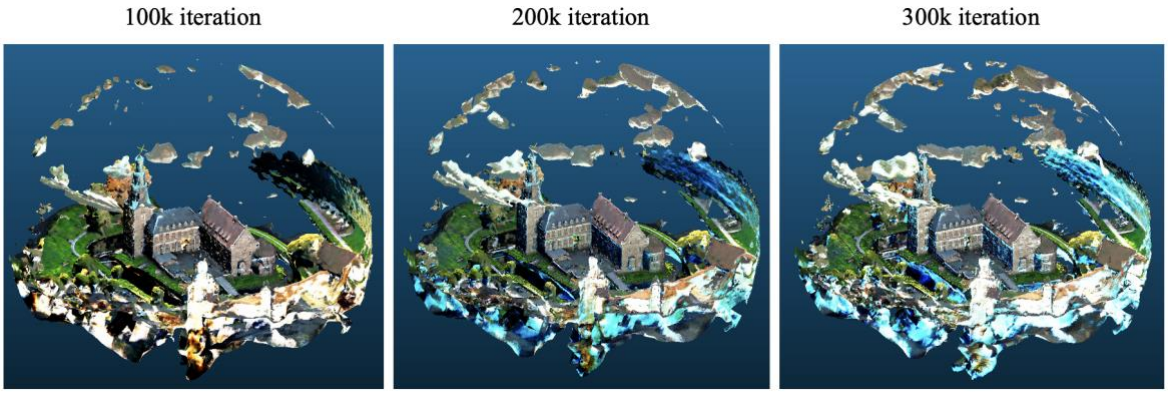

Fig. 7 Evolution of Neuralangelo surface reconstruction at different training iterations.

2) Final model selection

   Following Li et al. (2023), which identifies 300k iterations as a standard configuration for large-scale outdoor scenes, and consistent with observed geometric stabilization, the 300k iteration model was selected as the final Neuralangelo reconstruction.

3) Mesh extraction (deterministic settings)

   Mesh extraction was performed using the provided `extract_mesh.py` script with Marching Cubes at a fixed resolution of 2048 (`resolution = 2048`, `block_res = 128`). Only the largest connected component was retained (`keep_lcc`).

4) Conversion to point clouds for cross-method evaluation

   Final meshes were uniformly sampled to generate 1,000,000 points, followed by voxel-based downsampling at 1 cm resolution, as illustrated in Fig. 8. This standardization ensures direct comparability with COLMAP and other learning-based reconstructions under unified evaluation metrics.

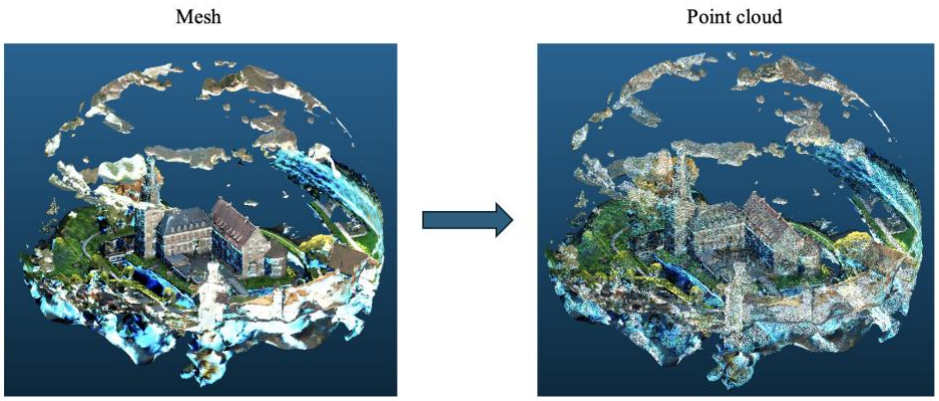

Fig. 8 Conversion of textured Neuralangelo mesh into point cloud representation.

4.3.2 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) is adopted in this study as a learning-based scene representation that departs from coordinate-based neural fields by explicitly modeling geometry using anisotropic 3D Gaussian primitives (Kerbl et al., 2023). All 3DGS reconstructions strictly follow the official implementation with default optimization parameters. In this formulation, a scene is represented as a collection of learnable Gaussians, each parameterized by a 3D mean position

$$\boldsymbol{\mu} \in \mathbb{R}^3,$$

an opacity value α, an anisotropic covariance matrix

$$\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3},$$

and view-dependent color coefficients encoded using spherical harmonics. This explicit geometric representation enables dense surface reconstruction while supporting efficient rendering and stable optimization for large-scale architectural scenes.

1. Camera Pre-processing, Undistortion, and Dataset Normalization

   Camera intrinsics and extrinsics required for 3DGS initialization were obtained through a COLMAP-based Structure-from-Motion (SfM) pre-processing stage, shared with the pipelines described in Sections 4.2.1. A SIMPLE_PINHOLE camera model was adopted to estimate camera poses and to generate an initial sparse point cloud, which served as the geometric seed for Gaussian initialization.

   The input dataset for 3DGS followed the canonical structure expected by the official implementation, consisting of:

   - `images/` containing undistorted input images, and

- `sparse/0/` containing COLMAP camera and pose parameters (`cameras.bin`, `images.bin`, `points3D.bin`).

By enforcing this standardized input contract, all reconstruction methods evaluated in this study operate on an identical camera solution, ensuring that observed differences arise from the reconstruction model rather than camera estimation.

2. Implicit Scene Representation and Optimization Assumptions

   In 3DGS, scene geometry is not represented as an explicit surface but emerges implicitly from the spatial distribution of overlapping Gaussian primitives. Each Gaussian defines a local density function in 3D space, and the rendered scene appearance is obtained by volumetric accumulation along camera rays.

   Given fixed camera poses, optimization proceeds by minimizing a photometric rendering loss between rendered pixel colors $\hat{C}(\mathbf{r})$ and observed image colors $C(\mathbf{r})$ along rays $\mathbf{r}$:

$$\mathcal{L}_{\text{photo}} = \sum_{\mathbf{r}} \left\| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 .$$

   No explicit surface regularization, signed-distance constraint, or depth supervision is imposed. Instead, geometric structure is encoded implicitly through the spatial arrangement, anisotropy, and opacity of the Gaussian set. During training, 3DGS alternates between parameter refinement and adaptive density control, in which Gaussians are cloned, split, or pruned based on reconstruction quality and visibility statistics (Kerbl et al., 2023).

3. Training Schedule, Run-time Configuration, and Reproducibility

   To examine the influence of training duration on reconstruction quality, two iteration settings were adopted: 7,000 and 30,000 iterations, as illustrated in Fig. 9, following the experimental design proposed by Kerbl et al. (2023). All 3DGS training was executed on PALMA High-Performance Computing (HPC) GPU nodes using NVIDIA A100 GPUs under headless execution conditions.

   For reproducibility, effective run-time arguments were retrieved from intact output directories generated by the official implementation. The training code automatically exports all command-line arguments to a `cfg_args` file in the output directory, which serves as the authoritative record of the effective configuration, analogous to the exported `config.yaml` used by Neuralangelo.

   The exported configuration specifies, among others:

- spherical harmonics degree $\mathrm{sh\_degree} = 3$,
- image resolution downsampling factor $\mathrm{resolution} = 2$,
- white background rendering assumption,
- GPU-based execution (`data_device = cuda`), and
- training from scratch without pretrained weights.

This explicit run-time traceability ensures that the 3DGS pipeline can be independently reproduced without reliance on manual intervention.

Fig. 9 Visual comparison of 3D Gaussian Splatting reconstructions at 7k and 30k iterations.

4. Splat-to-Point Cloud Conversion

While 3DGS produces a splat-based scene representation optimized for differentiable rendering, this native output is not directly compatible with conventional geometric evaluation pipelines. To enable quantitative comparison with other methods, all optimized 3DGS scenes were converted into explicit point cloud representations using a customized 3DGS-to-PC pipeline (Stuart & Pound, 2025), as illustrated in Fig. 10.

Fig. 10 Conversion of 3D Gaussian Splatting splat into point cloud representation.

Rather than extracting only Gaussian center points, this conversion samples points probabilistically from each Gaussian's 3D density distribution. For a Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, candidate points $\mathbf{x}$ are sampled such that their Mahalanobis distance satisfies:

$$(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \sigma^2,$$

with $\sigma = 2.0$. This constraint restricts sampling to statistically meaningful regions of each Gaussian's density field and suppresses extreme outliers.

The extraction process was configured to generate approximately 20 million points per scene, ensuring sufficient spatial density for architectural analysis. Color values were assigned based on the dominant view-dependent contributions of each Gaussian across all camera rays. The resulting point clouds were further refined using statistical outlier removal implemented in the Open3D library (Zhou et al., 2018), yielding clean, geometrically consistent datasets suitable for alignment, completeness analysis, and feature-based evaluation.

4.3.3 VGGT

Visual Geometry Grounded Transformer (VGGT) is adopted in this study as a pretrained learning-based multi-view geometry inference pipeline for estimating camera geometry and dense depth from image collections (Wang et al., 2025). In contrast to COLMAP, which reconstructs explicit structure via feature matching, triangulation, and bundle adjustment (BA), VGGT primarily operates through feed-forward neural inference and optionally performs a BA-based refinement stage. The VGGT workflow is reported as an executable pipeline, including (i) pretrained model specification and inference assumptions, (ii) view-subset protocol and input contract, (iii) feed-forward geometry prediction and failure-aware BA refinement, and (iv) depth-to-point-cloud export for cross-method evaluation.

1. Pretrained Model Usage and Inference Assumptions

   All VGGT experiments were conducted using the official implementation with a fixed pretrained checkpoint loaded via:

   $$\theta^{*} = \texttt{VGGT.from\_pretrained("facebook/VGGT-1B")}.$$

   The model was used as-is for inference only; no scene-specific fine-tuning or retraining was performed.

   Inference was executed on PALMA High-Performance Computing (HPC) GPU nodes using NVIDIA A100 GPUs, with mixed-precision acceleration enabled

(bfloat16). To ensure stable execution and reproducibility in the HPC environment, all model and dependency caches (TORCH_HOME, HF_HOME, XDG_CACHE_HOME) were redirected to scratch storage.

2. View-subset Protocol and Input Contract

   To evaluate scalability and sensitivity to input image count, each site was reconstructed using controlled subsets of increasing size:

   $$N \in \{8, 12, 24, 36, 48\}.$$

   For each subset, input images were organized under a standardized directory structure containing only an `images/` folder, which constitutes the canonical input contract expected by the official VGGT demo pipeline. No prior sparse reconstruction or external camera estimation was required.

   Camera geometry was inferred internally by VGGT under a SIMPLE_PINHOLE camera assumption, with no explicit modeling of lens distortion. This controlled subset design enables direct attribution of reconstruction stability and degradation to view-count scaling rather than uncontrolled dataset variation.

3. Feed-forward Geometry Prediction and Failure-aware BA Refinement

   Given a set of T input images, VGGT jointly predicts:

   - camera extrinsics $\mathbf{E}_t \in \mathbb{R}^{4\times4}$,
   - camera intrinsics $\mathbf{K}_t \in \mathbb{R}^{3\times3}$,
   - and per-view dense depth maps $\mathbf{D}_t \in \mathbb{R}^{H\times W}$.

   Depth prediction is conditioned on globally aggregated multi-view features produced by a transformer-based aggregator. For each camera ray r, depth is inferred directly without enforcing explicit surface regularization, signed-distance constraints, or volumetric consistency. As a result, geometric coherence emerges implicitly from learned feature correspondences rather than from an explicit geometric optimization objective.

   Consequently, VGGT's geometric consistency is not only encoded within the model architecture itself, but is conditioned on the spatial distribution and overlap characteristics of the input views. Empirically, overly redundant image sets with high overlap tend to degrade reconstruction stability, whereas geometry-aware sampling strategies such as farthest point sampling (FPS) better align with the model's implicit assumptions by promoting global view diversity rather than local redundancy.

VGGT optionally supports a BA refinement stage that attempts to improve camera poses and sparse geometry by minimizing reprojection error over inferred feature tracks. In this study, BA was enabled when sufficient geometric support was available, using the official demo entry point with fixed runtime parameters (e.g., `seed = 42`, SIMPLE_PINHOLE camera model, `max_query_pts = 2048`, and `query_frame_num = min(`N`, 5)`). All effective arguments were printed at runtime and recorded in per-subset log files, ensuring full configuration traceability.

However, BA refinement is not guaranteed to succeed, particularly as the number of input views increases. Two principal failure modes were observed:

1) Insufficient inlier support.

   For larger subsets (e.g., Schloss Münster with $N = 36$), the inferred feature tracks did not yield enough geometrically consistent inliers, leading to early BA termination ("Not enough inliers per frame"). This indicates that, beyond a certain view count, the learned correspondences produced by VGGT may fail to maintain sufficient multi-view consistency to support classical geometric refinement.

2) GPU memory exhaustion.

   For dense subsets (e.g., Schloss Raesfeld with $N = 48$), the fine-tracking and correlation-volume computation required by BA exceeded available GPU memory, resulting in CUDA out-of-memory errors. This behavior reflects the unfavorable scaling of the fine-grained tracking stage with respect to image count and feature resolution.

These failure modes explain the observed degradation of VGGT performance for larger image sets and highlight intrinsic scaling constraints of the current implementation.

In preliminary experiments, image subsets of 8, 12, 24, 36, 48, 72, 96, and 144 images were evaluated. However, reconstructions produced using 144 images consistently exhibited degraded geometric quality, characterized by increased noise and reduced structural coherence, even when BA was disabled or bypassed. In addition, recent studies such as OmniVGGT report that optimal VGGT performance is typically achieved within relatively small input ranges, often between 2 and 24 images (Peng et al., 2025). Based on these empirical observations and the identified failure modes at higher view counts, this study adopts three representative dataset

configurations—8, 24, and 48 images—to capture VGGT behavior under low, medium, and relatively high view-count regimes, as illustrated in Fig. 11.

Fig. 11 Comparison of VGGT reconstructions under different input image configurations.

Rather than enforcing BA as a mandatory stage, VGGT was integrated into a failure-aware pipeline. When BA refinement failed due to insufficient inliers or GPU memory constraints, the pipeline automatically reverted to feed-forward inference results without interrupting execution.

This design ensures that:

- BA is treated as an optional refinement, not a prerequisite.
- All image subsets yield a valid geometric output.
- Reconstruction failures do not propagate into downstream evaluation stages.

4. Depth Unprojection and Point Cloud Generation

To enable quantitative comparison with other reconstruction methods, VGGT outputs were converted into explicit point clouds via depth unprojection. For a pixel $\mathbf{u} = (u, v)$ with predicted depth $D_t(\mathbf{u})$, the camera-space point is computed as:

$$\mathbf{X}_c = D_t(\mathbf{u})\, \mathbf{K}_t^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix},$$

and transformed into world coordinates via:

$$\mathbf{X}_w = \mathbf{E}_t \begin{bmatrix} \mathbf{X}_c \\ 1 \end{bmatrix}.$$

RGB values were assigned by resizing the input images to the depth-map resolution and sampling per-pixel colors. The point clouds produced by VGGT do not natively include color information. To enable visual inspection and consistency with other reconstruction methods evaluated in this study, color attributes were

assigned in a post-processing stage using a colored point cloud approach implemented in Open3D (Zhou et al., 2018), as illustrated in Fig. 12.

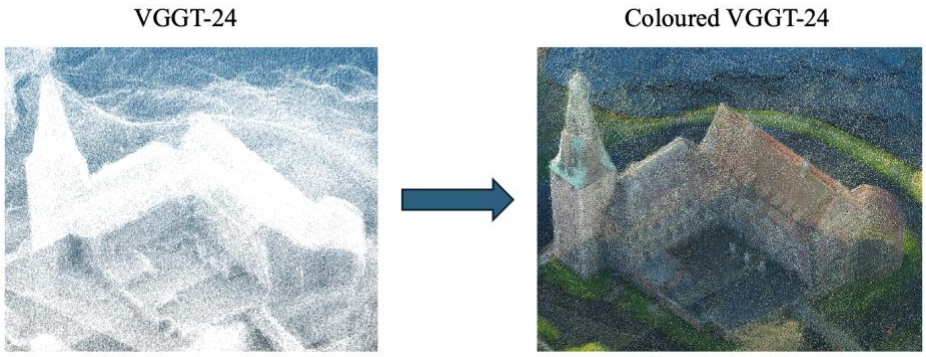

Fig. 12 Colorization of VGGT point clouds.

Overall, VGGT is treated in this study not as a fully optimized reconstruction pipeline but as a pretrained geometry inference system with fixed internal assumptions. Its performance is bounded by the representational capacity and scaling behavior learned during pretraining, rather than by scene-specific optimization or explicit geometric regularization.

4.4 Point Cloud Refinement

Raw point clouds generated by traditional photogrammetry (COLMAP) and learning-based reconstruction methods (Neuralangelo, 3D Gaussian Splatting, and VGGT) contain geometric noise and outliers. To ensure a fair and consistent basis for geometric comparison across Schloss Münster (M), Schloss Raesfeld (R), and Burg Lüdinghausen (L), a systematic point cloud refinement pipeline was applied using CloudCompare.

Point cloud refinement is a critical step in cultural heritage documentation, as filtering strategies directly influence geometric accuracy, surface continuity, and the interpretability of architectural features (Chen et al., 2021). Given the heterogeneous nature of the reconstruction outputs, refinement procedures were adapted to the geometric characteristics of each method rather than enforcing a single uniform filter configuration. While automatic filters such as Statistical Outlier Removal (SOR) and radius-based filtering are effective for stochastic noise, manual cleaning remains essential for removing large-scale artifacts, disconnected components, and environmental clutter commonly present in UAV-based heritage datasets (Bieńkowski & Rutkowski, 2022).

3D Gaussian Splatting (3DGS)

3DGS reconstructions frequently exhibit peripheral “floater” artifacts arising from splat-based density propagation. After manual removal of disconnected components, a noise filter

with a radius of 0.05 m was applied to suppress isolated points. Subsequently, SOR filtering ($k = 6, n\sigma = 2.0$) was used to reduce residual noise while preserving sharp architectural edges (Yu et al., 2025). The refined point clouds were then spatially resampled to a uniform resolution of 0.01 m to standardize density for comparison.

VGGT

VGGT outputs show substantial variation in point density and noise characteristics depending on the number of input images (8, 24, and 48). To account for this variability, radius-based filtering was adaptively configured, with finer radii (0.002–0.003 m) applied to most subsets and a coarser radius (0.04 m) used for sparse 8-image reconstructions of R and L. For large point sets, Octree-based subsampling (Level 10) was employed to manage data volume while preserving structural coherence, a strategy well suited to architectural heritage datasets (Bieńkowski & Rutkowski, 2022). For the 48-image subsets, a finer spatial resampling of 0.002 m was applied to retain detailed geometric features.

Neuralangelo and COLMAP

Neuralangelo reconstructions, benefiting from SDF-based regularization, exhibited relatively smooth and coherent surfaces and required minimal filtering beyond manual cleaning and uniform spatial resampling at 0.01 m. In contrast, COLMAP reconstructions generated through traditional MVS required more aggressive statistical filtering to address high-frequency noise. SOR filtering ($k = 8, n\sigma = 1.5$) was applied to M and L, and radius-based filtering was calibrated to site-specific scale differences to maintain geometric consistency (Demantke et al., 2011).

Across all methods, final spatial subsampling was applied to ensure comparable point densities prior to geometric alignment and evaluation. As emphasized by Demantke et al. (2011), appropriate scale selection is essential for capturing architectural dimensionality. By adopting scale-aware refinement strategies, this study ensures that subsequent quantitative evaluation reflects reconstruction behavior rather than artifacts introduced by inconsistent point cloud densities, as illustrated in Fig. 13.

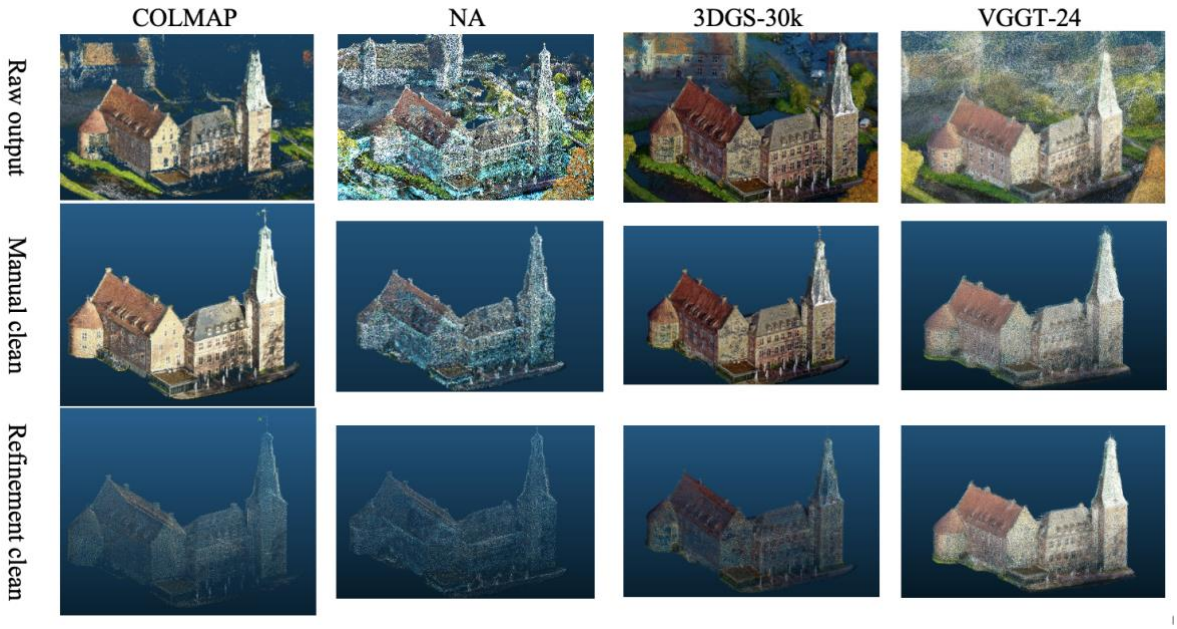

Fig. 13 Effect of point cloud refinement from COLMAP, Neuralangelo, 3DGS-30k, and VGGT-24.

4.5 Geometric Alignment and Registration

To conduct a rigorous geometric comparison between the different reconstruction outputs (COLMAP, Neuralangelo, 3DGS, and VGGT) and the ground-truth LiDAR data, all models must be transformed into a unified coordinate system. This study implements a two-stage registration workflow, initial manual alignment followed by Iterative Closest Point (ICP) refinement, to eliminate discrepancies in scale, orientation, and translation (Beraldin, 2004; Gaong et al., 2025).

Firstly, coarse registration was conducted using the Align (point pairs picking) tool. For each dataset, 13–15 homologous point pairs were manually selected between the reconstructed model and the LiDAR reference. Point selection was guided by the presence of clearly identifiable geometric features shared by both datasets (Weinmann, 2016). The selected points were spatially distributed across the entire structure to provide robust geometric constraints and reduce local bias. During this stage, scale adjustment was enabled to compensate for the inherent scale ambiguity present in neural reconstruction methods. The resulting transformation served as the baseline alignment for all subsequent analyses, as illustrated in Fig. 14.

Fig. 14 Manual alignment between reconstructed model and LiDAR reference.

After manual point-based alignment, ICP refinement was applied to further minimize the Euclidean distance between corresponding surfaces (Lague et al., 2013). ICP was configured with 80 iterations, a final overlap ratio of 70% to accommodate partial occlusions, and a random sampling limit of 300,000 points. To reduce the influence of extreme outliers, the farthest point removal option was enabled. Alignment quality was evaluated using a dual criterion combining quantitative and qualitative assessment. First, the Root Mean Square (RMS) distance reported by ICP was compared against the RMS of the manual alignment. Second, the resulting alignment was visually inspected to assess façade flushness, roof continuity, and edge consistency with the LiDAR reference. If ICP produced both a lower RMS value and an improved visual correspondence, the ICP-refined transformation was adopted for subsequent completeness and geometric feature analysis. However, in cases where ICP introduced geometric skewing—typically due to non-uniform point distributions or incomplete LiDAR coverage—the manually aligned result was retained, even if the RMS value was lower.

This hybrid alignment strategy ensures that registration accuracy is not assessed by numerical metrics, but also by geometric plausibility, supporting a fair and interpretable comparison across reconstruction methods and sites.

4.6 Evaluation Metrics

To assess the performance of different UAV-based 3D reconstruction methods, this study employs a consistent evaluation procedure focusing on three complementary aspects: accuracy, completeness, and geometric feature characteristics. All evaluations are conducted using the same software and analysis workflow in CloudCompare.

Accuracy is assessed by measuring the spatial deviation between reconstructed point clouds and the LiDAR reference after geometric registration. Alignment results obtained from manual correspondence-based registration and ICP refinement are both considered, and accuracy is quantified using Root Mean Square (RMS) distance metrics (Beraldin, 2004). Because LiDAR coverage varies across sites, alignment quality is interpreted in conjunction with visual inspection to ensure that low RMS values correspond to meaningful geometric correspondence rather than partial or biased overlap (Gaong et al., 2025). This combined interpretation allows accuracy to be evaluated consistently across reconstruction methods and heritage sites.

Completeness is assessed using cloud-to-cloud (C2C) distance analysis between the refined reconstructed point cloud and the LiDAR reference. Unlike ICP, which is employed for geometric alignment, C2C distance analysis operates on already aligned point clouds and measures spatial deviations on a per-point basis, serving purely as an evaluation metric. After computing C2C distances, reconstructed points are classified according to their distance to the LiDAR reference (Clini et al., 2024). Points within a distance threshold of 0–1 m are considered geometrically consistent with the reference data, while points exceeding this threshold indicate larger deviations. Completeness is then quantified as the ratio of reconstructed points within the 0–1 m range to the total number of points in the refined model (Lague et al., 2013). By focusing on distance-based correspondence rather than point density, this definition enables fair comparison across reconstruction methods that produce point clouds with substantially different densities and structural characteristics.

Geometric feature analysis is conducted to examine how reconstruction methods represent local surface properties (Weinmann, 2016). Nine geometric descriptors—roughness, curvature, surface density, omnivariance, eigenentropy, anisotropy, planarity, linearity, and sphericity—are computed to characterize local geometric behavior.

By applying the same evaluation procedures and parameter settings across all reconstruction results, this evaluation strategy supports consistent and transparent comparison. The combined use of accuracy, completeness, and geometric feature analysis provides a balanced assessment of reconstruction performance and forms the basis for the analyses presented in Chapter 5.

5 Analysis and Results

This chapter presents the analysis and results of the comparative evaluation of UAV-based 3D reconstruction methods. To provide a structure assessment, the analysis is divided into qualitative and quantitative perspectives, reflecting visual reconstruction quality and measurable geometric performance. Given that the three case-study castles differ substantially in architectural geometry, and LiDAR reference coverage, results are first examined on a per-site basis before conducting cross-castle comparisons.

For clarity, Schloss Münster, Burg Lüdinghausen, and Schloss Raesfeld are hereafter referred to as M, L, and R, respectively. Neuralangelo is denoted as NA, while the two 3D Gaussian Splatting configurations are referred to as 3DGS-7k and 3DGS-30k. VGGT reconstructions using 8, 24, and 48 input images are denoted as VGGT-8, VGGT-24, and VGGT-48, respectively. These notations are used consistently throughout Chapter 5.

5.1 Visual Qualitative Comparison

The visual qualitative comparison aims to examine the perceptual and structural characteristics of reconstructed models produced by different reconstruction frameworks. Reconstruction methods are grouped into two methodological categories: traditional photogrammetric reconstruction, represented by COLMAP, and learning-based reconstruction approaches, represented by Neuralangelo (NA), 3D Gaussian Splatting (3DGS), and VGGT. To account for configuration-dependent behavior within learning-based methods, 3DGS is evaluated using two training regimes (7k and 30k iterations), and VGGT is evaluated using three input image configurations (8, 24, and 48 images). This design enables an assessment of how training intensity and input data quantity influence visual reconstruction characteristics.

Visual appearance is analysed separately for each castle to account for differences in architectural geometry and LiDAR data completeness. Reconstructions generated by the same method may exhibit distinct visual characteristics across sites, reflecting the interaction between method design, dataset configuration, and scene geometry. The qualitative analysis adopts site-specific visual perspectives tailored to local architectural features and reference data availability. This approach allows meaningful site-dependent reconstruction patterns to be identified and prevents oversimplified visual comparison under heterogeneous heritage documentation conditions.

5.1.1 Visual Evaluation Criteria

This section focuses on evaluating how different reconstruction methods perform within the same castle, with an emphasis on visual quality. The analysis examines the visual appearance of the reconstructed models across the three castles, considering multiple aspects that influence reconstruction outcomes. One of the primary factors is the completeness of the available LiDAR reference data. Given the distinct architectural geometries of the three castles and the varying completeness of their LiDAR reference data, reconstructions generated by the same method may exhibit different visual characteristics under different site conditions. Therefore, site-specific visual evaluation criteria are defined for each castle, rather than enforcing a single unified standard across all sites.

Schloss Münster (M)

With respect to the LiDAR reference data, the dataset for M exhibits incomplete point cloud coverage along the side façades, while the roof structure is relatively well captured. As a result, during geometric alignment, only roof regions can be consistently aligned across all reconstruction methods. This constraint is nevertheless applied uniformly, ensuring fair comparison conditions. Consequently, quantitative evaluation for M is restricted to roof-level geometry, whereas visual evaluation allows for broader inspection of structural characteristics beyond the aligned regions.

For visual assessment prior to refinement, two primary architectural features are considered: (1) the central tower structure and (2) the large frontal façade. These features provide clear geometric and visual cues for comparing reconstruction quality across methods, particularly in terms of structural continuity, point density, surface smoothness, and preservation of architectural detail.

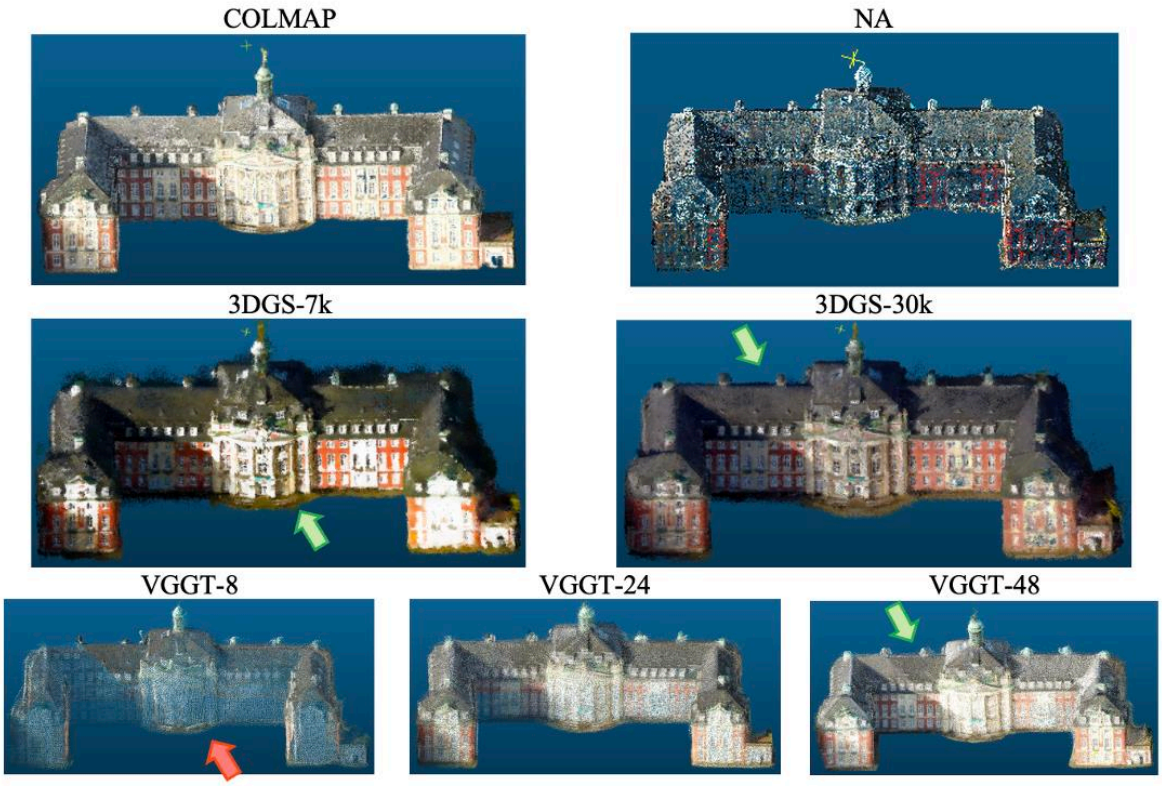

Fig. 15 Visual comparison for Schloss Münster across different reconstruction methods.

Under these visual evaluations shown in Fig. 15, the frontal façade of M is most sharply reconstructed by 3DGS-7k. The 3DGS-30k also performs well, though with slightly smoother surface characteristics. NA and VGGT-48 preserve recognizable façade structures, albeit with reduced sharpness and local detail compared to 3DGS. In contrast, COLMAP, VGGT-8 and VGGT-24 show weaker preservation of façade details, with visibly reduced texture definition and geometric clarity.

Among all evaluated configurations, VGGT-8 produces the sparsest reconstruction for M. Compared to other reconstruction methods, this model exhibits lower point density and fragmented surface representation, particularly along façade regions.

Differences are also observed in roof smoothness and overall structural continuity. Reconstructions produced by 3DGS-30k and VGGT-48 tend to generate more coherent roof surfaces, whereas lower-data or lower-iteration configurations exhibit increased surface irregularities and incomplete geometric transitions.

Overall, the visual comparison for M indicates that reconstruction quality is strongly influenced by point density and surface representation strategies, with 3DGS achieving superior façade detail preservation despite the absence of façade-level reference data. This highlights the importance of qualitative inspection for sites with incomplete LiDAR

coverage, where visual fidelity cannot be fully captured by roof-level quantitative evaluation alone.

Burg Lüdinghausen (L)

Similar to M, the LiDAR reference data for L provide relatively complete coverage of the roof surfaces, while the side façades lack sufficient point cloud information. Consequently, during geometric alignment, only the roof regions can be consistently aligned across all reconstruction methods. This limitation applies to all methods and is treated as a shared limitation of the comparative setup. However, the restriction to roof-level LiDAR reference data may bias both alignment and quantitative metrics toward roof-level agreement, potentially underrepresenting discrepancies in façade or fine-grained architectural details. As with M, quantitative evaluation for L is restricted to roof-level geometry, whereas visual assessment allows for a broader inspection of architectural details.

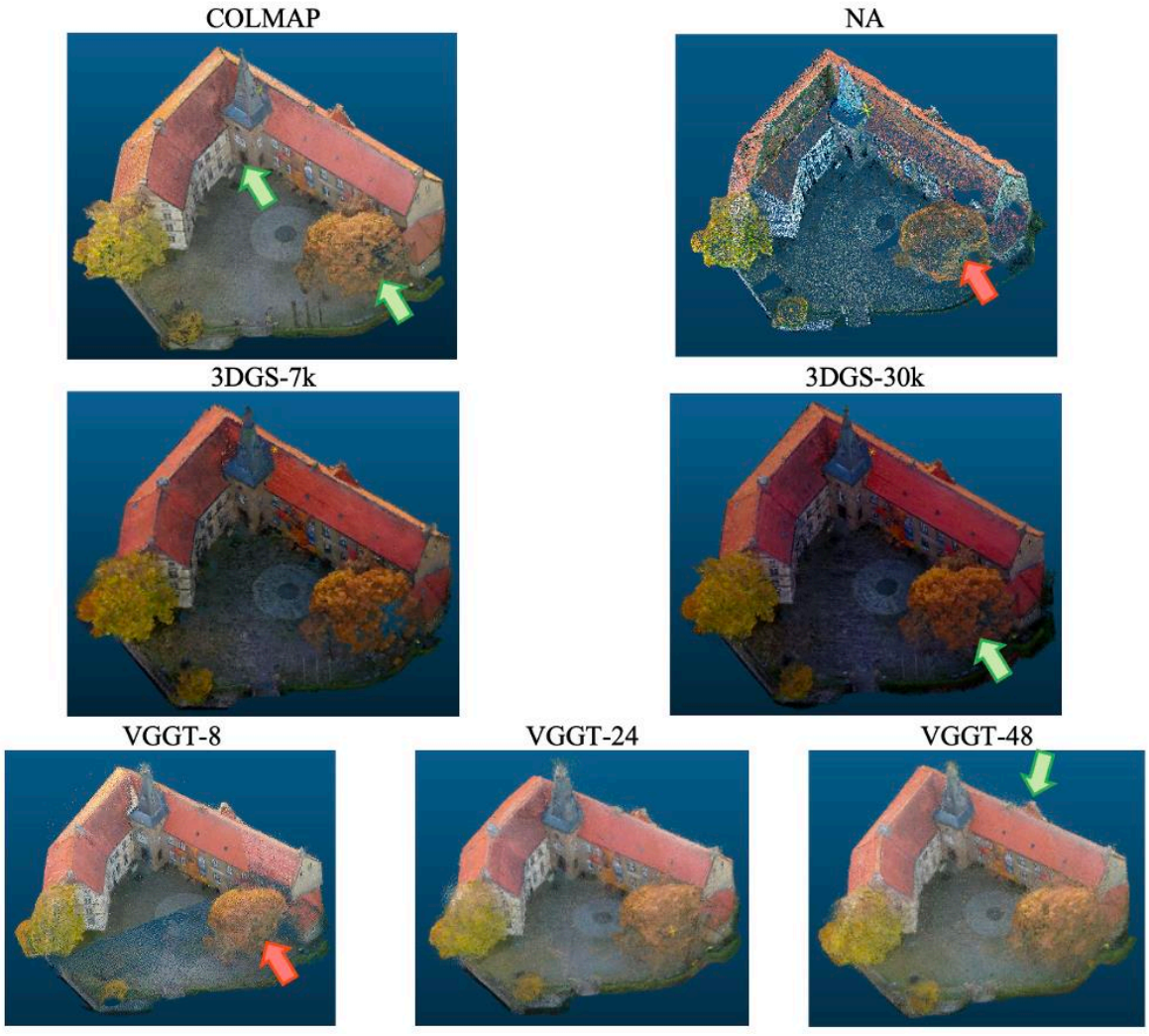

Fig. 16 Visual comparison for Burg Lüdinghausen across different reconstruction methods.

In contrast to M, the visual evaluation shown in Fig. 16 for L emphasizes finer architectural openings and the interaction between built structures and surrounding vegetation. Several site-specific architectural features are particularly informative for evaluating reconstruction quality, including the small windows on the central tower, the white-framed windows along the side façades, and the overall façade articulation of the central tower. Although these façade elements are not covered by the LiDAR reference data and are excluded from quantitative alignment and evaluation, they are visually well defined in the reconstructed models and provide valuable cues for qualitative comparison. In addition, a protruding dormer window on the roof represents a distinct architectural element that is visually separable from the main roof surface and serves as a useful reference for assessing reconstruction fidelity.

A site-specific challenge for L arises from the data acquisition conditions. UAV imagery was collected during the autumn season, resulting in the inclusion of two dense trees located in the castle courtyard. While these trees are partially present in the LiDAR reference data, the corresponding point clouds are considerably less complete than those reconstructed from UAV imagery. As a result, during alignment, the reconstructed tree geometries exhibit noticeable discrepancies relative to the LiDAR data. This mismatch may appear as reduced completeness in quantitative evaluation; however, it reflects limitations in reference data coverage rather than deficiencies in the reconstruction itself, analogous to the absence of LiDAR coverage for the castle façades.

Overall, the visual reconstruction quality for L remains relatively stable and coherent across the evaluated methods. None of the methods produces models that are excessively sparse or overly compact, and the main architectural structures are consistently preserved. Notable differences are primarily observed in the reconstruction of vegetation within the courtyard. COLMAP and both 3DGS configurations reconstruct the tree structures with relatively high completeness, preserving branching patterns and foliage extent. In contrast, the NA tends to represent the trees as dense, aggregated foliage masses with less clearly distinguishable branching structures. VGGT-8 appears more simplified in tree geometry compared to the 24- and 48-, which capture more detailed tree shapes and spatial extent.

Beyond vegetation, subtle differences are also visible in roof surface continuity and edge definition. Higher-data configurations, such as 3DGS-30k and VGGT-24 or 48-, tend to produce smoother roof surfaces and more coherent transitions between roof segments, whereas lower-data configurations show slightly increased surface irregularities. Taken together, L represents a reference-aligned but visually informative site, where quantitative

metrics are less sensitive to method differences, and qualitative evaluation reveals method-specific behaviour in vegetation handling and surface continuity rather than in primary architectural reconstruction.

Schloss Raesfeld (R)

R is the site with the most complete LiDAR reference data among the three castles, with both roof surfaces and façade regions well covered. As a result, visual evaluation for this site allows for a more comprehensive inspection of reconstruction quality compared to M and L. A notable discrepancy between the reconstructed models and the LiDAR reference data is the presence of a row of parasols in the reconstructed scenes, which are not captured in the LiDAR data. Since this discrepancy occurs consistently across all reconstruction methods, the resulting misalignment affects all methods equally and does not introduce bias in the comparative analysis. This discrepancy is treated as a site-specific artifact rather than a factor influencing method performance.

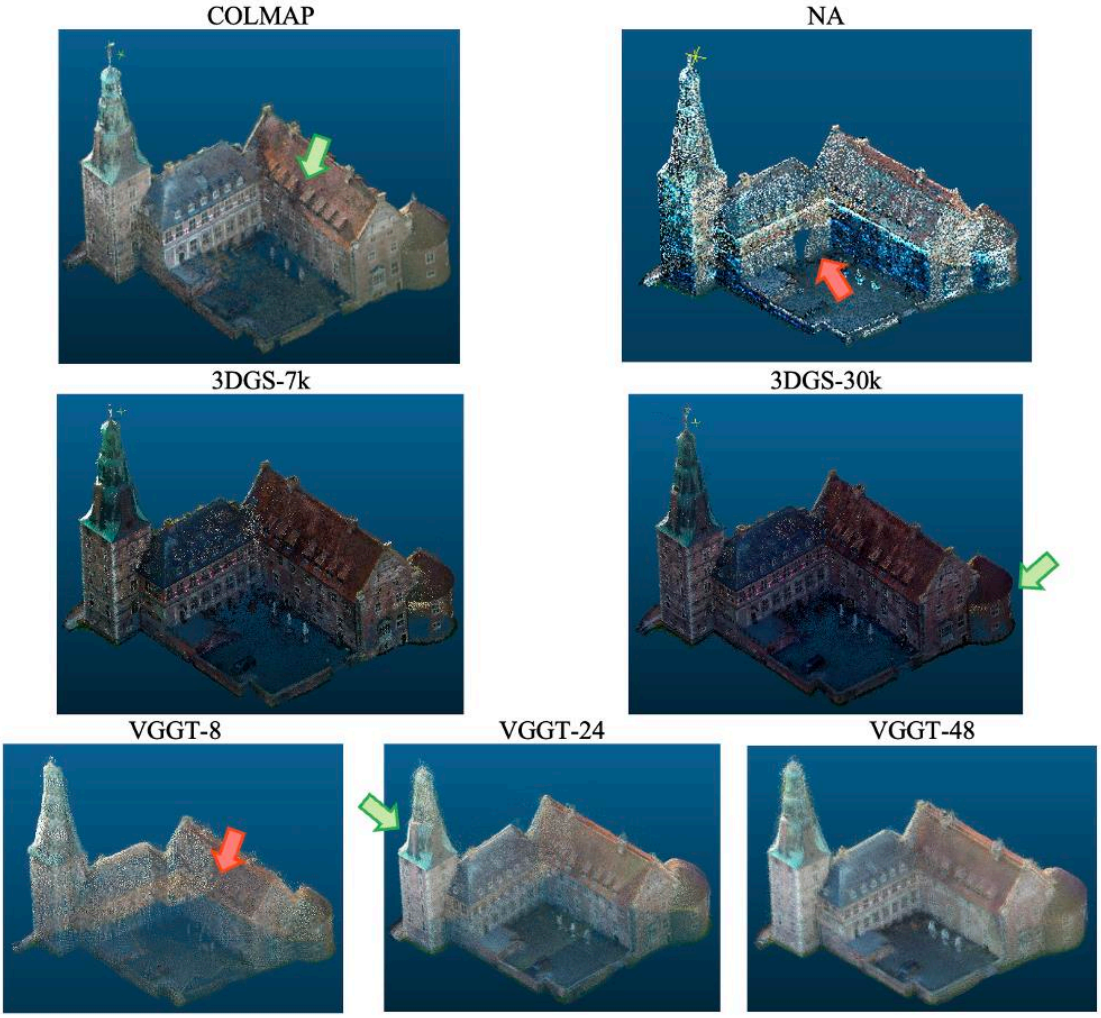

Fig. 17 Visual comparison for Schloss Raesfeld across different reconstruction methods.

For visual evaluation shown in Fig. 17, several architectural features are informative for R. One key feature is the pronounced bending along the side of the tower. Reconstruction methods may differ in their sensitivity to angular changes, and insufficient detection of this bending can lead to overly smoothed or simplified geometry, making this feature suitable as a visual evaluation criterion. Another important feature is the circular structure connected to the main castle body. This structure is characterized by a circular roof and vertically extending walls, and the continuity between the roof and wall surfaces provides a clear indication of reconstruction completeness. Additionally, a row of outward-protruding windows represents a strongly 3D architectural element. The extent to which these windows are reconstructed with clear volumetric definition serves as another useful indicator of geometric fidelity.

Among the three sites, R exhibits the most pronounced visual differences across reconstruction methods. The VGGT reconstructions reveal a clear trend across the three dataset configurations. With VGGT-8, the reconstructed model preserves the overall structural outline of the castle, but fine details appear blurred and simplified. Nevertheless, the model does not exhibit strongly underrepresented or excessively sparse regions. VGGT-24 becomes noticeably denser and more coherent across the entire structure, without excessive concentration in specific regions. VGGT-48 enhances the visibility of surface appearance, in terms of color clarity; however, the additional geometric improvement relative to 24- is limited.

In contrast, NA performs noticeably worse for R compared to the other reconstruction methods. While the overall external outline of the castle remains recognizable, finer architectural details are poorly preserved and appear substantially smoothed or indistinct. COLMAP and both 3DGS configurations show clearer preservation of architectural details, including sharper roof edges and more consistent façade structures. Between the two 3DGS configurations, increasing iterations results in smoother surfaces and improved structural continuity, while the overall geometric layout remains similar.

Overall, R provides the most informative setting for distinguishing method-specific reconstruction behaviour due to the completeness of its LiDAR reference data and the richness of its architectural features. This makes R particularly suitable for highlighting method-specific sensitivities to architectural complexity and reference data completeness. Accordingly, R functions as a control site in this study, where comprehensive reference coverage enables clear discrimination of method-specific reconstruction behaviour under minimal reference-induced bias.

5.1.2 Cross-Castle Visual Qualitative Performance

This section focuses on cross-castle visual comparison, examining how the same reconstruction method performs across different castles.

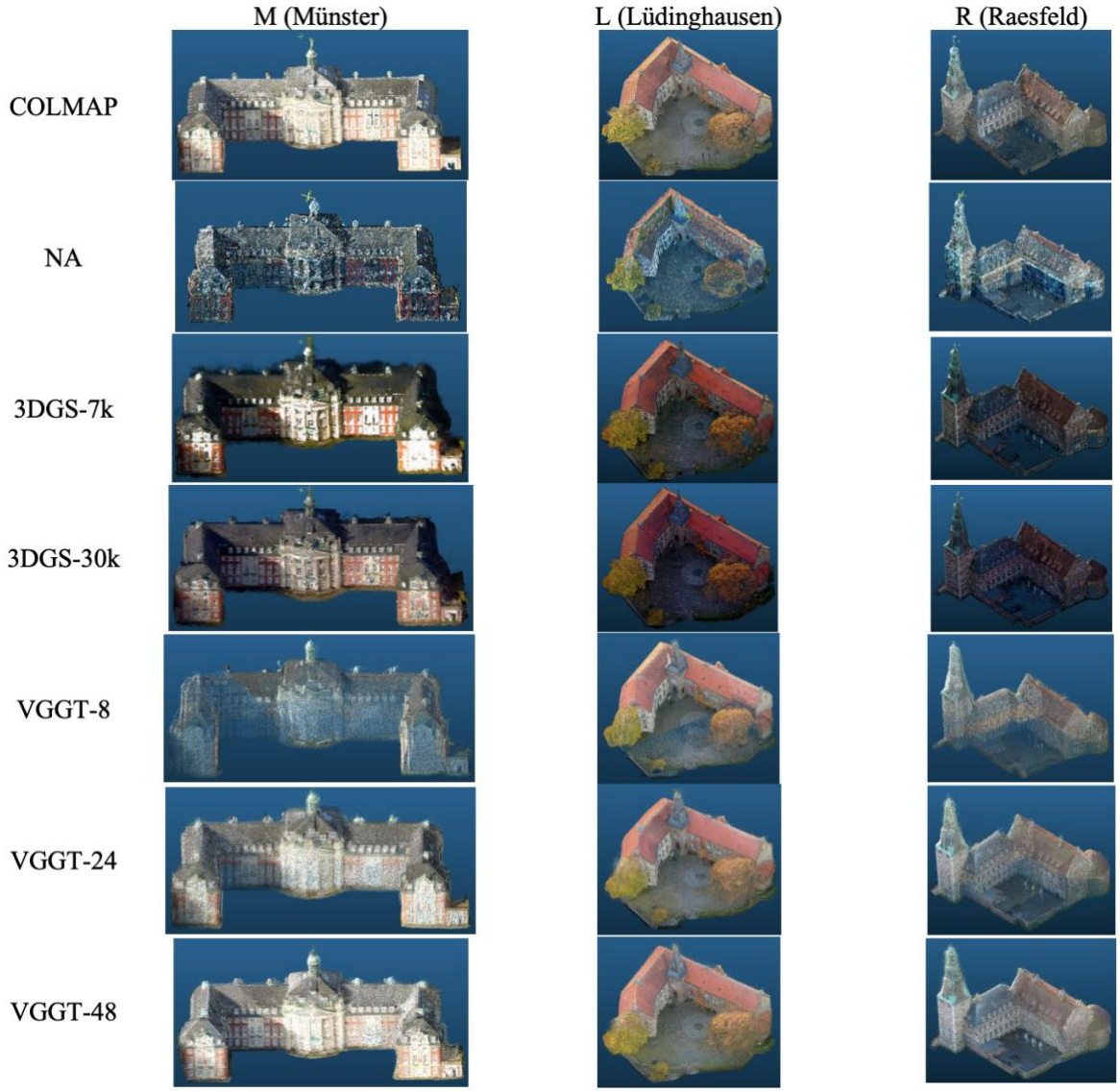

Fig. 18 Visual comparison for cross-castle across reconstruction methods and configurations.

Based on the cross-castle visual comparison shown in Fig. 18, COLMAP across the three castles, and excluding variations caused by lighting conditions during data acquisition, the method demonstrates consistently strong performance in terms of overall geometry, structural detail, completeness, and surface fidelity. One limitation of COLMAP is its strong dependence on the original image data, as the method does not perform additional color optimization during reconstruction. As a result, exposure-related artifacts present in the input imagery are directly propagated into the reconstructed models. This characteristic makes COLMAP highly sensitive to acquisition conditions, as manual parameter adjustment during

reconstruction offers limited ability to correct such effects. Nevertheless, in terms of geometric reconstruction, COLMAP remains one of the most stable and reliable methods across all three castles.

NA consistently produces models with a smaller overall spatial extent compared to the other reconstruction methods. To visually inspect the NA reconstructions in detail, magnification is often required. After refinement, NA also yields the lowest number of points among all methods, indicating that the reconstructed geometry is compact and evenly distributed. In contrast, other methods may exhibit localized gaps when aggressive point cloud cleaning is applied. Since NA produces texture-based surface reconstructions, the models in this study are converted to point clouds using a mesh-to-point-cloud pipeline (Zhou et al., 2018). This pipeline is implemented using a custom Python script based on Trimesh and Open3D, with parameters set to one million uniformly sampled points and voxel-based uniform sampling with 1 cm spacing. This conversion process may also influence the resulting visual appearance of the NA reconstructions. NA exhibits comparatively weaker reconstruction quality across the three castles. While the overall outlines of the cultural heritage structures remain recognizable, finer elements are generally blurred. For R, the reconstruction is largely limited to coarse structural contours, with architectural features such as windows and tower bending poorly preserved. For M, window locations can still be identified, although fine details remain indistinct. Among the three sites, L exhibits the most complete NA reconstruction, with the overall building form and some architectural details more clearly represented.

An interesting observation emerges when comparing the two 3DGS configurations across the three castles. Increasing the number of training iterations from 7k to 30k does not result in substantial changes to the overall geometric structure. The 7k configuration already produces stable geometric reconstructions for all sites, indicating that further increases in training iterations are not strictly necessary from a geometric perspective. Differences in geometric detail are generally subtle for L and R, which exhibit fewer fine-grained architectural features. In contrast, for M, which presents higher geometric complexity, the 30k configuration leads to more clearly reconstructed details, such as more distinct window structures on the right-hand façade and more pronounced roof ridge lines along the lateral extensions of the main building. While geometric differences remain limited, increasing the number of training iterations primarily affects visual appearance. Across all three castles, the 30k configuration consistently exhibits darker color tones compared to the 7k results, despite comparable point densities. These color shifts are particularly noticeable for M,

where roof surfaces transition to a different color range. Overall, 3DGS demonstrates stable visual reconstruction quality across all three sites, with higher training iterations mainly enhancing surface appearance and local visual clarity.

Notable differences are observed among the three VGGT dataset configurations. Visually, VGGT-8 produces a noticeably sparser reconstruction for M compared to its performance on L and R. In contrast, VGGT reconstructions for L remain consistently complete across all three dataset configurations (8, 24, and 48 images), with stable overall structure and density. For R, VGGT exhibits a pattern similar to M; however, the reconstruction produced with 8- is less sparse than the corresponding result for M. Overall, no strong monotonic trend is observed across the three VGGT datasets when all castles are considered together.

The most pronounced cross-castle variation is observed in the VGGT-8. Under this condition, M and R exhibit noticeably sparser reconstructions compared to L. Apart from reduced ground-level detail in L, the overall visual reconstruction for this site does not appear sparse. This behavior is closely related to scene geometry under extremely sparse input conditions. VGGT-8 relies heavily on learned correspondences derived from a very limited number of views, making reconstruction quality strongly dependent on whether the selected images adequately capture the dominant geometric variations of the scene.

M is characterized by a long, horizontally extended façade with strong central symmetry and substantial depth variation along the façade. With VGGT-8, this combination of horizontal extent and façade depth cannot be sufficiently captured, resulting in stable reconstruction only in limited regions and a visually fragmented overall structure. In contrast, L exhibits a more compact geometry with continuous roof surfaces and an enclosed courtyard, forming favorable conditions for high view overlap even with a small number of images. As a result, VGGT-8 performs reliably for this site, producing a visually dense and coherent reconstruction. R, characterized by a tall tower, bent wall structures, and strong vertical variation, presents a different challenge: eight images are insufficient to simultaneously capture both vertical structures and wall bending, leading to reconstructions that convey the overall structure but remain partially incomplete.

These observations indicate that under extremely sparse input conditions, the visual reconstruction quality of VGGT is strongly influenced by scene geometry. In contrast, VGGT-24 and VGGT-48 exhibit highly consistent overall shape, density, and structural continuity across all three castles. Overall, the cross-castle visual analysis reveals that reconstruction robustness across sites is method-dependent. While COLMAP and 3DGS maintain stable geometric and visual characteristics across all three castles, VGGT exhibits

pronounced sensitivity to scene geometry under extremely sparse input conditions. This sensitivity largely diminishes as the number of input images increases, with VGGT-24 and VGGT-48 converging toward consistent visual quality across sites. These results highlight scene geometry as a critical factor governing cross-site generalization for learning-based inference methods under limited-view regimes.

5.2 Quantitative Performance Evaluation

The quantitative performance evaluation is conducted from three complementary perspectives: accuracy, completeness, and geometric features. In addition, a cross-castle quantitative comparison is performed to examine how these metrics vary across different sites. Alongside numerical analysis, visual inspections of completeness are employed as supportive qualitative references to facilitate interpretation of the quantitative results.

5.2.1 Accuracy Comparison

Across the three castles, alignment accuracy is governed by the interaction between reconstruction method and site characteristics. Schloss Münster (M) exhibits mixed alignment behaviour due to geometric complexity and incomplete LiDAR coverage. Burg Lüdinghausen (L) consistently benefits from ICP refinement, while Schloss Raesfeld (R), supported by the most complete reference data, demonstrates clear performance differentiation among reconstruction methods, favouring traditional photogrammetry-based approaches.

Schloss Münster (M)

For M, alignment accuracy exhibits relatively unstable behaviour across different reconstruction methods. As shown in Table 1, RMS values range from 0.55 to 1.50, depending on the reconstruction method and the alignment strategy. COLMAP, NA, and 3DGS-30k achieve lower RMS values when using manual point selection, whereas 3DGS-7k and all VGGT configurations (8, 24, and 48 images) achieve lower RMS values after ICP refinement.

This bimodal behaviour indicates that no single alignment strategy consistently outperforms the other for M. One possible explanation lies in the geometric configuration of the site. M is characterized by a long, horizontally extended façade with strong symmetry and limited LiDAR coverage on the lateral façades. Manual point selection may benefit methods that preserve clear and visually identifiable architectural features on the main façade, while ICP

alignment may favour reconstructions that produce more uniformly distributed point clouds over roof surfaces, which dominate the available LiDAR reference data. As a result, alignment performance for M appears to be strongly influenced by the interaction between reconstruction characteristics and site geometry.

Table 1 RMS errors for Schloss Münster across different reconstruction methods

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
RMS (Manually)	0.96	1.03	1.50	1.37	1.16	1.33	1.16
RMS (ICP)	1.15	1.17	1.33	1.46	0.72	**0.55**	0.95

Burg Lüdinghausen (L)

In contrast to M, L exhibits a more consistent alignment behaviour across reconstruction methods. As indicated in Table 2, ICP refinement consistently reduces RMS values for all evaluated methods, leading to markedly improved alignment accuracy. Methods that show relatively high RMS values under manual alignment—such as 3DGS-30k and VGGT-48—experience substantial error reduction after ICP refinement.

A particularly notable result is observed for VGGT-24, which achieves the lowest RMS value among all methods after ICP refinement. This suggests that, for L, VGGT-24 may provide a favourable balance between point distribution and geometric coherence, enabling ICP to converge towards a stable and accurate alignment. Overall, these results demonstrate that ICP refinement is highly effective for L, yielding consistently precise alignment across different reconstruction approaches.

Table 2 RMS errors for Burg Lüdinghausen across different reconstruction methods

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
RMS (Manually)	1.41	2.33	1.70	3.57	1.90	2.22	3.96
RMS (ICP)	1.11	1.35	1.13	1.13	1.14	**0.70**	1.01

Schloss Raesfeld (R)

For R, ICP refinement leads to a clear and consistent improvement in alignment accuracy across all reconstruction methods. As shown in Table 3, RMS values after ICP refinement are lower than those obtained through manual alignment. R represents the site with the most complete LiDAR reference data among the three case studies, including comprehensive roof and façade coverage, which provides favourable conditions for robust global alignment.

Even prior to ICP refinement, COLMAP and 3DGS-30k already achieve relatively low RMS values using manual point selection. Although VGGT-8 and VGGT-48 exhibit higher RMS errors under manual alignment, these errors are effectively reduced after ICP refinement.

After ICP refinement, a clear performance ordering emerges for R, with COLMAP achieving the lowest RMS values, followed by NA, then 3DGS (7k and 30k), and finally VGGT. This ordering suggests that, for sites with complete reference data and well-defined geometry, traditional photogrammetry methods tend to achieve higher geometric accuracy than learning-based approaches.

Overall, while ICP refinement improves alignment accuracy for all evaluated methods, COLMAP consistently attains the lowest RMS errors for R, indicating a persistent advantage in geometric accuracy under favourable alignment conditions.

Table 3 RMS errors for Schloss Raesfeld across different reconstruction methods

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
RMS (Manually)	0.53	1.49	2.53	0.79	3.02	2.64	3.82
RMS (ICP)	**0.28**	0.34	0.38	0.40	0.46	0.44	0.48

Taken together, the accuracy comparison demonstrates that alignment performance is strongly site-dependent and closely linked to the completeness and spatial distribution of the LiDAR reference data. While ICP refinement consistently improves RMS errors across all reconstruction methods, its effectiveness varies by site: alignment behaviour remains unstable for M due to limited façade reference coverage, becomes highly robust for L, and yields clear method-level performance ordering for R. Under favourable reference conditions, traditional photogrammetry (COLMAP) achieves the highest geometric accuracy, whereas learning-based methods exhibit greater sensitivity to alignment strategy and reference data characteristics.

5.2.2 Completeness Comparison

For the qualitative completeness comparison, one representative configuration is selected for each reconstruction framework to enable method-level interpretation. 3DGS-30k and VGGT-24 are used as representative settings for 3DGS and VGGT, respectively, based on their overall quantitative performance across sites, while COLMAP and NA are included directly as single-configuration methods. This strategy avoids confounding effects from internal parameter variations and ensures consistent qualitative comparison across castles.

Overall, completeness results exhibit strong site dependency. VGGT-24 consistently achieves relatively high completeness across all three castles, while Schloss Raesfeld (R) shows the most stable and uniformly high completeness values regardless of reconstruction method. In contrast, Schloss Münster (M) and Burg Lüdinghausen (L) demonstrate greater

sensitivity to reconstruction density and distance-based filtering, reflecting the influence of LiDAR coverage and architectural configuration on quantitative completeness evaluation. Taken together, the qualitative completeness results in Figures 18–20 show that completeness is not only a function of reconstruction density but is mediated by site-specific LiDAR coverage and architectural configuration. Learning-based methods tend to preserve broader spatial coverage within the accepted distance threshold, while photogrammetric methods exhibit completeness patterns more tightly coupled to reference geometry.

Schloss Münster (M)

For M, completeness is strongly constrained by limited LiDAR roof coverage, leading to a clear dependence on how reconstructed points spatially align with the reference rather than on overall reconstruction density. Methods that achieve lower RMS errors after alignment generally exhibit higher completeness values, as summarized in Table 4. Among all evaluated approaches, VGGT-24 attains the highest completeness (65.05%), followed by VGGT-8 (58.63%) and VGGT-48 (52.66%). In contrast, 3DGS-7k (38.14%) and 3DGS-30k (35.43%) yield substantially lower completeness values.

Table 4 Completeness metrics for Schloss Münster across different reconstruction methods.

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
Refined Cloud	123,791	46,426	402,850	503,498	138,828	528,587	161,875
Points 0–1m	64,471	22,638	153,635	178,384	81,401	343,857	85,242
Completeness (%)	52.08%	48.76%	38.14%	35.43%	58.63%	**65.05%**	52.66%

This behaviour can be partially attributed to the limited LiDAR coverage of M, where reliable reference data are primarily available for roof regions. Completeness is defined here as the proportion of reconstructed points within a 0–1 m distance threshold from the LiDAR reference. Methods that generate dense but spatially concentrated reconstructions, such as 3DGS, therefore tend to lose a larger fraction of points after distance-based filtering. In contrast, VGGT-24 preserves a higher proportion of points within the accepted distance threshold, despite differences in absolute point counts across methods.

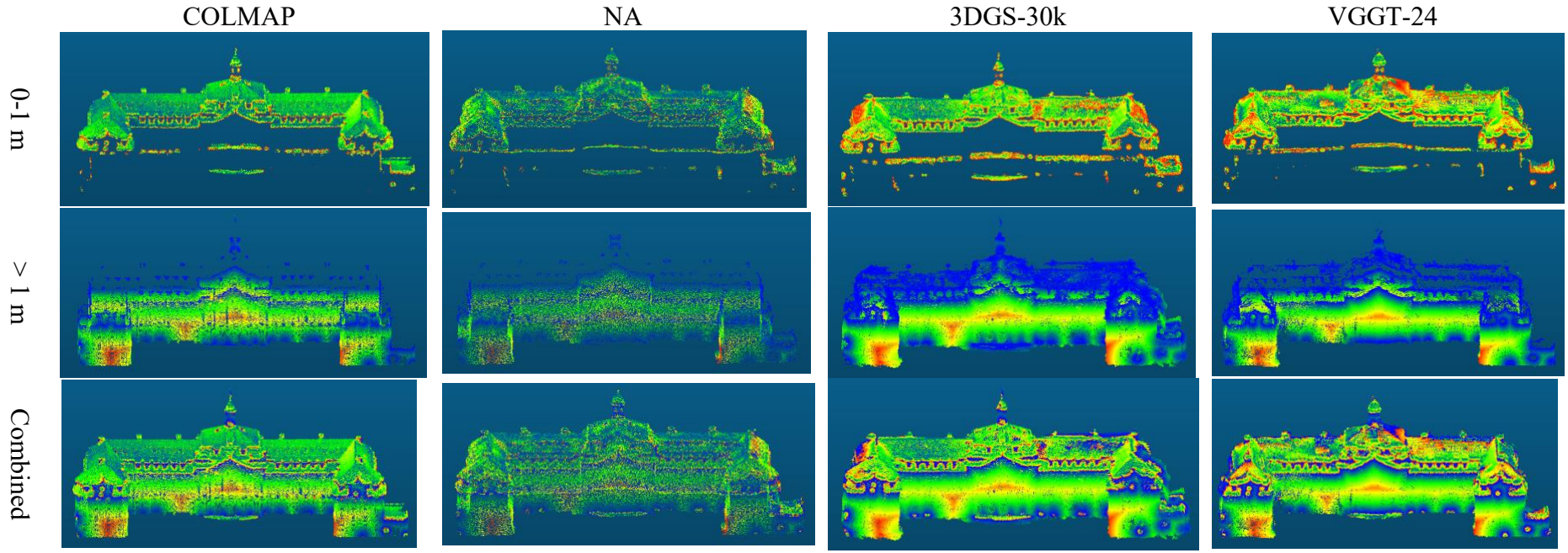

Fig. 19 Qualitative comparison of completeness for Schloss Münster across reconstruction methods.

In addition to the quantitative completeness values, the spatial distribution of cloud-to-cloud distances provides insight into how completeness should be interpreted for M, as illustrated in Fig. 19. For COLMAP and NA, points within the 0–1 m threshold are mainly concentrated on the roof surfaces captured by the LiDAR reference, resulting in moderate completeness values and limited spatial extent beyond the main structure. For 3DGS-30k, despite relatively low completeness values, the 0–1 m distance map shows a largely continuous reconstruction of the main castle geometry, while points exceeding the threshold are primarily located around façade boundaries and peripheral regions. In contrast, VGGT-24 achieves higher completeness through broader spatial coverage within the distance threshold.

Burg Lüdinghausen (L)

L represents a reference-aligned completeness regime, where quantitative completeness primarily reflects correspondence with architectural reference data rather than overall visual reconstruction quality. For L, completeness values are more evenly distributed across reconstruction methods, ranging from 40.31% to 59.94%, as shown in Table 5. Similar to M, VGGT-24 again achieves the highest completeness (59.94%), while NA records the lowest value (40.31%). The remaining methods, including COLMAP and both 3DGS variants, cluster within a relatively narrow range of approximately 45% to 47%, indicating reduced variability in completeness across methods.

Table 5 Completeness metrics for Burg Lüdinghausen across different reconstruction methods.

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
Refined Cloud	121,848	63,442	372,655	433,455	173,493	714,342	424,159
Points 0–1m	58,109	25,572	174,016	196,812	77,863	428,153	215,244
Completeness (%)	47.69%	40.31%	46.70%	45.41%	44.88%	**59.94%**	50.75%

Compared to the other sites, L exhibits a more geometrically compact structure with consistent roof coverage, which mitigates extreme differences in reconstruction completeness. As a result, L represents a comparatively stable and interpretable case for completeness evaluation, where method-dependent variations are less pronounced than those observed for sites with more complex geometry or uneven reference coverage.

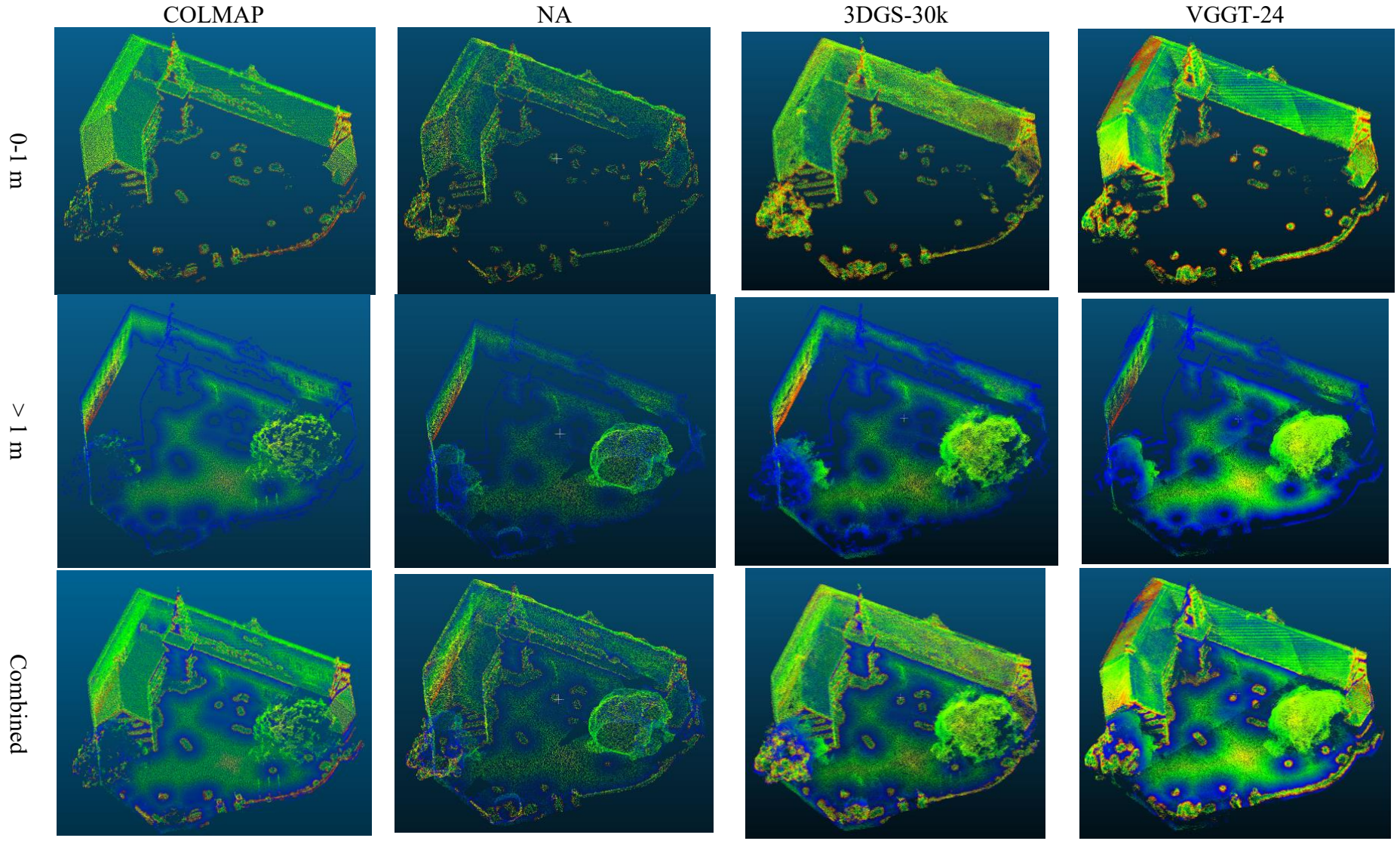

Fig. 20 Qualitative comparison of completeness for Burg Lüdinghausen across reconstruction methods.

The C2C distance visualization for L, as illustrated in Fig. 20, shows that all four methods—COLMAP, NA, 3DGS-30k, and VGGT-24—produce visually coherent reconstructions of the castle structure. Differences in quantitative completeness are mainly driven by non-architectural elements. 3DGS-30k and VGGT-24 reconstruct surrounding vegetation and ground surfaces with high visual quality; however, as these elements are not included in the LiDAR reference, they fall outside the distance threshold and reduce completeness values. In contrast, COLMAP and NA focus more strictly on the architectural structure, resulting in a closer correspondence with the reference data. This indicates that, for L, quantitative completeness reflects reference correspondence rather than overall visual reconstruction quality.

Schloss Raesfeld (R)

For R, completeness remains consistently high across all reconstruction methods, with values ranging from 75.63% to 83.65%, as shown in Table 6. The highest completeness is achieved by COLMAP (83.65%), followed closely by VGGT-24 (80.52%) and NA (79.81%). Even the lowest-performing configuration, VGGT-8, maintains a completeness level above 75%, indicating limited sensitivity of completeness to reconstruction method for this site.

Table 6 Completeness metrics for Schloss Raesfeld across different reconstruction methods.

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
Refined Cloud	72,579	42,089	160,325	181,057	147,670	651,288	264,782
Points 0–1m	60,712	33,592	125,902	139,130	111,676	524,436	207,434
Completeness (%)	**83.65%**	79.81%	78.53%	76.84%	75.63%	80.52%	78.34%

In contrast to M, no clear inverse or direct relationship between alignment accuracy and completeness is observed for R. Instead, both metrics remain consistently high across methods. This stability is likely attributable to the more complete LiDAR coverage of R, which includes roof and façade information, reducing the sensitivity of completeness measurements to reconstruction density and alignment strategy.

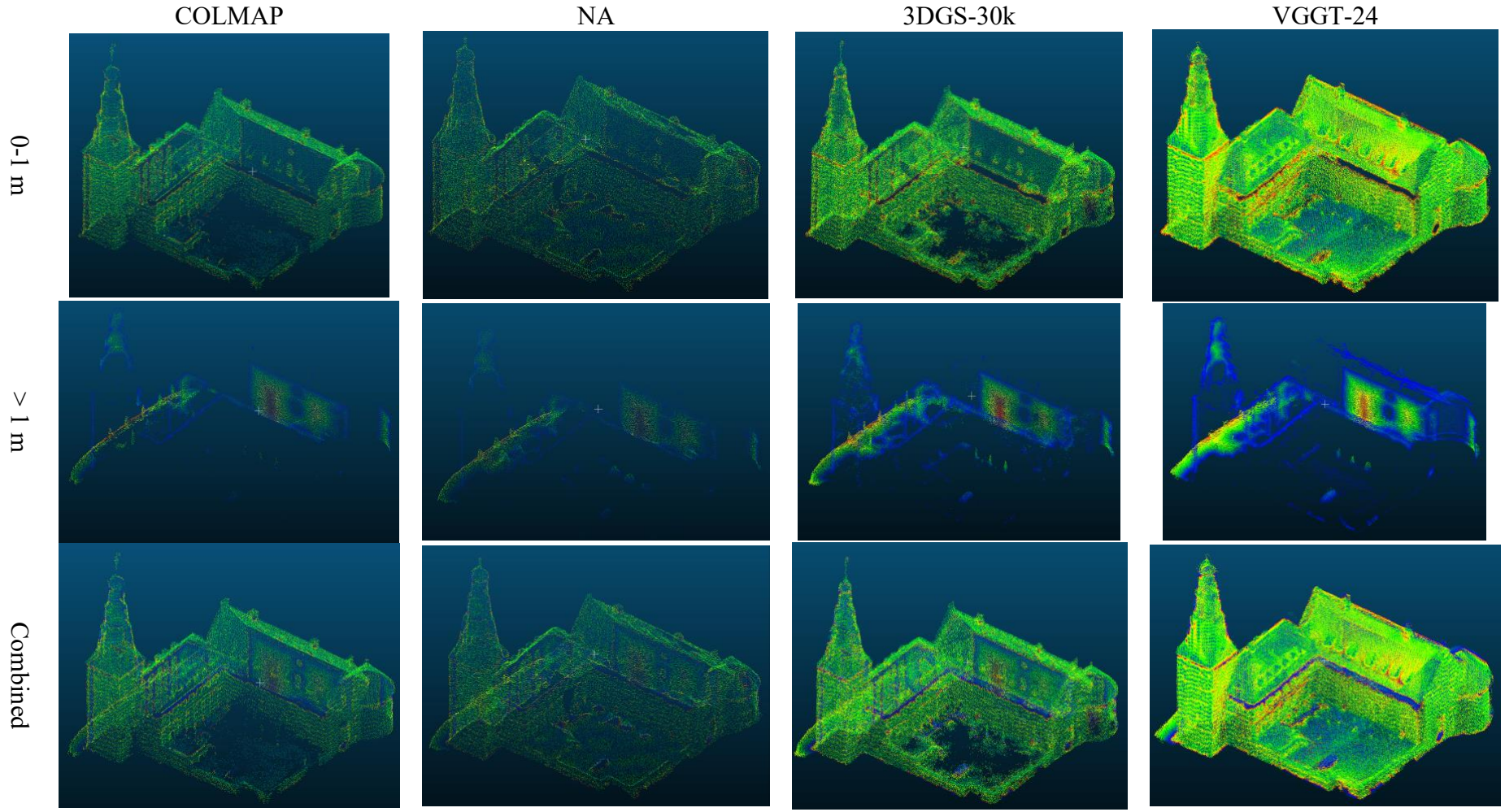

Fig. 21 Qualitative comparison of completeness for Schloss Raesfeld across reconstruction methods.

The qualitative C2C distance visualization for R, as illustrated in Fig. 21, supports the quantitative completeness results. Across all four methods—COLMAP, NA, 3DGS-30k, and VGGT-24—the majority of reconstructed points fall within the 0–1 m threshold and are consistently distributed over roof and façade surfaces. Unlike M and L, no systematic concentration of excluded points is observed in peripheral or non-architectural regions. This visual consistency confirms that, for R, high completeness values correspond to comprehensive reconstruction of the castle geometry rather than to reference-dependent filtering effects. As such, R serves as a control case in which high and uniformly distributed reference coverage minimizes method-dependent variability, allowing completeness to more reflect reconstruction fidelity.

5.2.3 Geometric Feature Analysis

Geometric feature analysis is performed using nine local geometric descriptors to characterise how different reconstruction methods represent surface structure. This analysis focuses on relative patterns in geometric behaviour across methods and sites.

Overall, the results indicate clear method-dependent behaviours. Traditional photogrammetry (COLMAP) consistently produces smoother, more planar, and directionally stable surfaces, whereas learning-based approaches tend to increase surface density, curvature sensitivity, and local irregularity. Among the VGGT configurations, VGGT-24 provides a balanced representation with relatively high density and moderate roughness and curvature across sites. Importantly, geometric feature responses are strongly influenced by site geometry, underscoring that these descriptors should be interpreted in a relative and site-aware manner.

Schloss Münster (M)

For M, geometric feature responses vary noticeably across reconstruction methods, as summarized in Table 7. Roughness values are generally low for COLMAP and NA, indicating smoother reconstructed surfaces, while 3DGS variants and VGGT-8 exhibit medium to high roughness, suggesting increased surface irregularity. VGGT-48 shows the highest roughness, reflecting denser but less regular local surface structure.

Table 7 Mean geometric feature values for Schloss Münster across different reconstruction methods.

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
Roughness	0.05	0.05	0.07	0.07	0.07	0.04	0.10
Curvature	0.23	0.14	1.33	1.87	0.41	1.16	0.95
Surface Density	13.05	4.62	17.81	15.74	17.17	41.09	8.34
Omnivariance	0.03	0.05	0.02	0.01	0.03	0.01	0.04
Eigenentropy	0.45	0.68	0.25	0.21	0.42	0.16	0.40
Anisotropy	0.96	0.98	0.69	0.78	0.88	0.83	0.75
Planarity	0.67	0.64	0.31	0.33	0.52	0.40	0.33
Linearity	0.29	0.35	0.38	0.45	0.35	0.43	0.42
Sphericity	0.04	0.02	0.31	0.22	0.12	0.17	0.25

In terms of curvature, COLMAP and NA remain in the low range, whereas 3DGS (especially 30k) and VGGT-24 show high curvature values, indicating stronger sensitivity to façade details and roof articulation. Surface density varies substantially, with VGGT-24 producing markedly higher densities compared to all other methods, while NA consistently yields low-density reconstructions.

Descriptors related to local shape complexity, such as omnivariance and eigenentropy, are generally low across all methods, though NA and COLMAP exhibit slightly higher eigenentropy, suggesting more heterogeneous local neighborhoods. Anisotropy and planarity are highest for COLMAP and NA, reflecting more planar and directionally consistent surfaces, while 3DGS and VGGT variants show reduced planarity, consistent with their denser but less strictly planar reconstructions. Sphericity remains low overall, with 3DGS variants showing relatively higher values, indicating a tendency toward more isotropic local point distributions.

Burg Lüdinghausen (L)

L exhibits more stable geometric feature distributions across reconstruction methods, as summarized in Table 8. Roughness values remain low to medium for all reconstructions, with minimal variation across methods, indicating generally smooth surface representations. Curvature, however, is notably higher for 3DGS and VGGT variants—particularly VGGT-24 and VGGT-48—reflecting stronger sensitivity to roof edges and architectural articulation within the compact building structure.

Table 8 Mean geometric feature values for Burg Lüdinghausen across different reconstruction methods.

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
Roughness	0.03	0.03	0.05	0.05	0.05	0.03	0.04
Curvature	0.28	0.19	1.62	1.94	0.58	1.64	1.69
Surface Density	22.66	9.80	34.12	32.35	23.06	104.86	39.88
Omnivariance	0.01	0.02	0.01	0.01	0.02	0.00	0.01
Eigenentropy	0.32	0.48	0.16	0.15	0.26	0.09	0.14
Anisotropy	0.96	0.98	0.68	0.72	0.86	0.79	0.72
Planarity	0.68	0.63	0.31	0.32	0.50	0.41	0.34
Linearity	0.28	0.35	0.37	0.39	0.36	0.38	0.38
Sphericity	0.04	0.02	0.32	0.28	0.14	0.21	0.28

Surface density shows the most pronounced variation at this site. VGGT-24 achieves substantially higher density than all other methods, while NA remains consistently low. This pattern aligns with the completeness analysis, where VGGT-24 also achieves the highest completeness, suggesting a consistent tendency toward dense local sampling. Omnivariance and eigenentropy remain low overall, indicating limited local shape variability across methods, although NA again exhibits slightly elevated eigenentropy.

High anisotropy and planarity values are observed for COLMAP and NA, reflecting their preference for planar and directionally consistent surface representations. In contrast, VGGT and 3DGS methods exhibit reduced planarity, consistent with their more flexible surface modeling strategies. Linearity and sphericity remain relatively stable across methods, suggesting that the compact and enclosed geometry of L constrains variability in these descriptors.

Schloss Raesfeld (R)

R presents the most consistent geometric feature patterns across reconstruction methods, as summarized in Table 9. Roughness remains low to medium, with 3DGS variants showing slightly elevated values compared to COLMAP and NA. Curvature is moderate overall, but increases noticeably for VGGT-24 and VGGT-48, reflecting enhanced representation of towers and curved wall segments.

Table 9 Mean geometric feature values for Schloss Raesfeld across different reconstruction methods.

	COLMAP	NA	3DGS-7k	3DGS-30k	VGGT-8	VGGT-24	VGGT-48
Roughness	0.06	0.06	0.10	0.09	0.07	0.04	0.06
Curvature	0.21	0.17	0.46	0.55	0.42	1.67	1.13
Surface Density	14.31	7.35	23.75	23.04	23.73	47.27	18.81
Omnivariance	0.04	0.05	0.04	0.03	0.03	0.01	0.02
Eigenentropy	0.56	0.69	0.38	0.36	0.38	0.13	0.27
Anisotropy	0.95	0.97	0.70	0.67	0.79	0.81	0.67
Planarity	0.70	0.71	0.42	0.38	0.50	0.37	0.32
Linearity	0.25	0.26	0.28	0.29	0.30	0.44	0.36
Sphericity	0.05	0.03	0.30	0.33	0.21	0.19	0.33

Surface density again differentiates methods clearly, with VGGT-24 producing the densest reconstructions, while NA yields the sparsest. Unlike the other sites, omnivariance and eigenentropy display marginally higher values for COLMAP and NA, suggesting that R's more complex façade geometry introduces greater local variation even in traditional reconstructions.

Anisotropy and planarity remain high for COLMAP and NA, indicating stable planar representations, while 3DGS and VGGT variants show reduced planarity, consistent with more volumetric point distributions. Sphericity values are generally low, though VGGT-48 exhibits a slight increase, indicating more isotropic local neighborhoods in denser reconstructions.

The geometric feature analysis reveals that reconstruction methods differ in how they encode local surface structure. Traditional photogrammetry consistently favours planar, anisotropic, and directionally stable surfaces, while learning-based methods trade strict planarity for increased surface density, curvature sensitivity, and volumetric richness. Importantly, absolute feature values are strongly modulated by site geometry and reference coverage, indicating that these descriptors are most meaningful when interpreted in a relative, site-aware manner rather than as absolute indicators of reconstruction quality. In this context, VGGT-24 emerges as a balanced configuration, achieving high local density while maintaining moderate roughness and curvature across all sites.

5.2.4 Cross-Castle Quantitative Performance

The cross-castle quantitative comparison provides a holistic perspective on reconstruction performance by jointly examining accuracy and completeness across Schloss Münster (M), Burg Lüdinghausen (L), and Schloss Raesfeld (R), as summarized in Table 10.

Table 10 Cross-castle comparison of accuracy (RMS) and completeness across reconstruction methods.

Method	M (Münster)		L (Lüdinghausen)		R (Raesfeld)	
	RMS	Comp.	RMS	Comp.	RMS	Comp.
COLMAP	0.96	52.08%	1.11	47.69%	**0.28**	**83.65%**
NA	1.03	48.76%	1.35	40.31%	0.34	79.81%
3DGS-7k	1.33	38.14%	1.13	46.70%	0.38	78.53%
3DGS-30k	1.37	35.43%	1.13	45.41%	0.40	76.84%
VGGT-8	0.72	58.63%	1.14	44.88%	0.46	75.63%
VGGT-24	**0.55**	**65.05%**	**0.70**	**59.94%**	0.44	80.52%
VGGT-48	0.95	52.66%	1.01	50.75%	0.48	78.34%

Across all methods, accuracy generally improves from M to L and reaches its best performance at R, as reflected by systematically lower RMS values at R. COLMAP demonstrates the most stable accuracy across all three castles, confirming the robustness of

traditional pipelines when sufficient geometric constraints and LiDAR coverage are available. Learning-based methods exhibit stronger site dependency. VGGT variants, particularly VGGT-24, achieve competitive accuracy in M and L but do not surpass COLMAP in R, where geometric completeness and alignment constraints are stronger. 3DGS variants show comparable accuracy between M and L but do not exhibit clear improvement with increased training iterations at the cross-castle scale.

Completeness reveals a pronounced site-dependent pattern. R consistently yields the highest completeness values across all methods, reflecting its more complete LiDAR coverage and well-defined architectural geometry. In contrast, M and L exhibit lower and more variable completeness, particularly for methods that produce dense but spatially concentrated point clouds. Among all approaches, VGGT-24 consistently achieves the highest completeness across sites, indicating strong robustness to site variation, whereas 3DGS configurations show systematically lower completeness in M and L, despite relatively dense reconstructions.

The relationship between accuracy and completeness is not consistent across sites. In M, lower RMS errors tend to coincide with higher completeness, whereas in R, both metrics remain uniformly high across methods without a clear trade-off. This suggests that the interaction between accuracy and completeness is strongly mediated by site geometry and LiDAR reference coverage.

Overall, COLMAP exhibits the most stable accuracy, while VGGT-24 emerges as the most robust method in terms of completeness. Learning-based approaches show increased sensitivity to site characteristics, particularly in geometrically extended or partially observed structures. These results indicate that no single method uniformly dominates across all quantitative criteria; instead, reconstruction performance must be interpreted in conjunction with site-specific geometry and reference data availability.

6 Discussion

This chapter discusses how reconstruction quality emerges from the interaction between scene geometry, UAV flight feasibility, data availability, evaluation workflow design, and the qualitative and quantitative findings.

6.1 Influence of Architecture and UAV Flight Conditions on Reconstruction Outcomes

A key observation of this study is that architectural geometry and surrounding environmental conditions influence the effectiveness of UAV-based reconstruction. Among the three castles, Schloss Raesfeld (R) consistently exhibits the most stable and high-quality reconstructions, followed by Burg Lüdinghausen (L), while Schloss Münster (M) shows the greatest variability in reconstruction quality across methods.

The instability observed for M can be attributed to the combination of two unfavorable conditions. First, dense surrounding vegetation—particularly due to the adjacent botanical garden—restricts UAV flight paths along the rear façade. In practice, this often required either increasing flight altitude to bypass vegetation or flying very close to the building envelope, both of which resulted in redundant image acquisition within limited battery capacity. Second, the large physical scale and strong bilateral symmetry of M pose additional challenges for reconstruction. Highly symmetric structures are inherently difficult for both traditional and learning-based methods, as repetitive façade patterns reduce the distinctiveness of feature correspondences. Under limited flight time and battery constraints, it becomes infeasible to capture sufficiently diverse viewpoints around such a large structure, a limitation that is clearly reflected in the sparse reconstruction obtained with VGGT-8.

In contrast, Burg Lüdinghausen represents the smallest structure among the three sites. While vegetation is located in close proximity to the building, UAV flight itself is relatively unconstrained. However, vegetation is frequently reconstructed together with architectural elements, and subsequent manual point cloud refinement introduces potential variability depending on how vegetation is removed. Despite this, the compact geometry of L allows for high viewpoint overlap even with a minimal number of images, explaining why VGGT-8 performs comparatively well at this site.

Schloss Raesfeld presents the most favorable conditions for UAV-based documentation. The castle is situated on open terrain without surrounding trees or taller neighboring structures, allowing consistent flight paths without the need for altitude adjustments. Moreover, the

moderate building size enables a greater number of viewing angles to be captured within the same battery budget. As a result, R consistently achieves the most complete and stable reconstructions across all evaluated methods.

6.2 Challenges in UAV Flight Strategy for Cultural Heritage Documentation

Cultural heritage structures present significantly greater flight planning challenges than regular urban buildings due to their complex geometry, irregular surroundings, and site-specific constraints. While established UAV flight strategies from the literature provide general guidance, effective data acquisition ultimately depends on the operator's situational judgment and experience.

The reconstruction results for Schloss Raesfeld illustrate this point clearly. The tower structure, including subtle variations in roof underside material and surface weathering, is consistently well reconstructed, indicating that the adopted flight strategy successfully captured critical viewpoints. Color consistency and geometric continuity in these regions further suggest that sufficient overlap and angular coverage were achieved.

At the same time, the results highlight that UAV-based cultural heritage documentation requires advanced piloting skills. Navigating terrain constraints, vegetation, and safety distances while maintaining optimal imaging geometry is non-trivial. Consequently, the effectiveness of reconstruction methods cannot be fully separated from the quality of UAV operation, making proficient flight execution a prerequisite for reliable downstream reconstruction.

6.3 Modeling Workflow and Data Representation Uncertainty

Beyond data acquisition, reconstruction outcomes are influenced by modeling workflow design and data representation choices. In this study, Neuralangelo reconstructions were converted from textured meshes to point clouds using a custom mesh-to-point-cloud pipeline, while 3D Gaussian Splatting outputs were similarly transformed using a separate custom script. Although consistent refinement strategies were applied thereafter, these intermediate conversions introduce additional sources of uncertainty, making it difficult to guarantee that comparisons strictly reflect the intrinsic behavior of the original reconstruction methods.

A similar concern arises for VGGT, which produces geometry without color information. Color attributes were subsequently assigned using Open3D, raising questions about whether such post-processing steps influence visual interpretation. These observations underscore

that cross-method comparison is not only affected by reconstruction algorithms themselves, but also by the representational transformations required to place heterogeneous outputs into a common evaluation space.

6.4 Point Cloud Refinement Strategy and Limitations

Point cloud refinement in this study followed literature-based guidelines to ensure methodological consistency. However, due to differences in building geometry and reconstruction characteristics, identifying a single refinement strategy applicable across all castles required iterative adjustment. While a common refinement pipeline was ultimately adopted, its parameters were necessarily defined at a coarse level to remain applicable across heterogeneous structures.

As a result, refinement quality was assessed not only based on point count reduction, but also in relation to structural completeness. This trade-off highlights the difficulty of defining universally comparable refinement templates and reinforces the need for flexible, site-aware processing strategies rather than rigid, fully standardized pipelines.

6.5 Geometric Alignment and Manual Correspondence Selection

Geometric alignment between reconstructed models and LiDAR reference data in this study was performed using manual correspondence selection, with approximately 13–15 control points per site. The number and spatial distribution of selected points were found to directly influence alignment outcomes. Moreover, the completeness of the reconstructed models and the LiDAR reference data constrained how many reliable correspondences could be identified. As a result, alignment strategies could not be uniformly defined, but instead had to be adapted to the specific conditions of each castle.

Because correspondence selection is performed manually, operator-dependent error is introduced into the alignment process. This represents an inherent limitation of the current workflow and constitutes an important source of uncertainty that should be acknowledged when interpreting quantitative accuracy and completeness metrics. The influence of human judgment in selecting alignment points is a relevant topic for discussion in comparative evaluation studies.

To reduce potential bias introduced by manual correspondence selection, an alternative strategy is to rely solely on Iterative Closest Point (ICP) alignment. However, observations from this study indicate that ICP alignment is highly sensitive to disparities in point cloud completeness and density. When substantial differences exist between the reconstructed

model and the LiDAR reference data, ICP may converge to a numerically favorable solution that is geometrically misleading.

This limitation is particularly evident in the alignment results for Schloss Münster (M). In this case, the LiDAR reference data provide complete coverage only for roof surfaces, while façade regions are largely missing. When ICP is applied directly under these conditions, the resulting alignment often yields lower RMS error values, suggesting improved numerical accuracy. However, visual inspection reveals noticeable spatial offsets between the reconstructed model and the LiDAR reference, indicating misalignment despite the reduced error metric. This discrepancy highlights that ICP optimization may favor regions with dense point correspondence (e.g., roofs) while neglecting large missing or unmatched areas, ultimately producing an alignment that is numerically stable but visually incorrect.

Based on these observations, ICP alignment appears suitable primarily for cases in which the reconstructed model and the reference point cloud exhibit comparable spatial extent and point density. When significant disparities exist—as in the case of incomplete LiDAR coverage—manual correspondence selection remains necessary to guide alignment, despite the introduction of operator-dependent uncertainty. This trade-off underscores the challenge of achieving objectivity and robustness in geometric alignment for cultural heritage sites and reinforces the importance of combining quantitative metrics with visual validation when interpreting alignment quality.

6.6 Interpretation of Method-Specific Performance

The comparative analysis reveals distinct performance profiles across traditional photogrammetry, neural scene representations, and geometry-grounded learning-based methods. COLMAP consistently demonstrates strong accuracy across all three castles, particularly in Schloss Raesfeld. Its stable performance can be attributed to the explicit geometric constraints of the photogrammetry pipeline, which directly optimize camera poses and point geometry. However, COLMAP's reliance on the original photometric conditions limits its ability to compensate for exposure inconsistencies or incomplete viewpoints, making its performance sensitive to data acquisition quality.

Neuralangelo (NA) exhibits a different behavior profile. While its reconstructions preserve global structure and overall shape continuity, fine-scale geometric details are often smoothed or underrepresented. This characteristic is reflected in lower completeness values and reduced geometric feature variability, particularly in complex architectural elements such as window protrusions and curved surfaces. NA's texture-driven reconstruction strategy,

combined with mesh-to-point cloud conversion, prioritizes surface continuity over local geometric fidelity.

3D Gaussian Splatting (3DGS) produces visually dense and continuous reconstructions with strong color consistency. However, increasing training iterations from 7k to 30k does not yield substantial improvements in geometric accuracy or completeness across sites. Instead, iteration increases primarily affect visual appearance, such as color saturation and contrast. This suggests that, for cultural heritage structures with well-defined geometry, 3DGS reaches geometric stability relatively early, and further training mainly refines appearance rather than structure.

VGGT demonstrates the most pronounced sensitivity to input image quantity. While VGGT-8 highlights the limitations of extremely sparse views, particularly for elongated or vertically complex structures, VGGT-24 consistently achieves a favorable balance between accuracy, completeness, and geometric stability across all castles. Increasing the number of images to 48 does not yield proportional quantitative gains, suggesting diminishing returns once sufficient geometric coverage is achieved. This behavior reflects a fundamental difference between optimization-based reconstruction pipelines and feed-forward geometry-grounded inference, where geometric consistency is learned implicitly rather than enforced through iterative error minimization.

6.7 Reproducible Comparative Framework

Beyond individual method performance, this study contributes a reproducible comparative framework for evaluating UAV-based 3D reconstruction methods in cultural heritage. By standardizing preprocessing, point cloud refinement, alignment strategies, and metric computation across all reconstruction pipelines, the framework ensures that observed differences primarily reflect reconstruction behavior. The explicit separation of qualitative visual assessment and quantitative metrics, together with site-aware interpretation, enables meaningful comparison even under heterogeneous LiDAR coverage conditions. Furthermore, the adoption of mean geometric feature categorization mitigates the risk of over-interpreting absolute metric values across geometrically distinct sites. As a result, the proposed framework provides a structured and transferable basis for comparative analysis, supporting future benchmarking efforts and promoting transparent and reproducible evaluation in cultural heritage 3D reconstruction research. The design of this framework also facilitates future reproducibility, as outlined in Chapter 8.

7 Conclusion

This thesis presented a comprehensive comparative analysis of traditional and learning-based UAV-based 3D reconstruction methods for cultural heritage applications. By evaluating COLMAP, Neuralangelo, 3D Gaussian Splatting, and VGGT across three architecturally distinct castles under real-world UAV acquisition conditions, the study examined reconstruction performance from visual, quantitative, and geometric perspectives. The results demonstrate that traditional photogrammetry remains highly reliable in terms of geometric accuracy, particularly when reference data coverage is sufficient. Learning-based methods introduce greater flexibility in handling sparse or irregular viewpoints; however, their performance is strongly influenced by input data characteristics, site geometry, and reconstruction configuration. Among the evaluated approaches, VGGT with a moderate image count (24 images) consistently achieved a balanced trade-off between accuracy, completeness, and geometric stability.

Importantly, the findings highlight that reconstruction performance cannot be reduced to a single metric or method ranking. Instead, meaningful evaluation requires a holistic, site-aware perspective that accounts for architectural geometry, UAV acquisition constraints, reference data completeness, and reconstruction objectives. The reproducible comparative evaluation framework proposed in this study provides a structured basis for such analysis and supports transparent and interpretable comparison across heterogeneous heritage sites.

Future work may extend this framework to additional heritage typologies, incorporate semantic or material-aware analysis, and explore hybrid reconstruction pipelines that combine the geometric robustness of traditional photogrammetry with the representational flexibility of learning-based methods.

8 Data Availability

To support future reproducibility, documentation of the comparative evaluation framework will be released at: https://github.com/tingjia-guo/HERI3D

The repository includes structured descriptions of preprocessing steps, reconstruction configurations, alignment procedures, and evaluation metrics used across all reconstruction pipelines, enabling transparent, consistent comparisons.

Bibliographical References

Ali, M. S., Zhang, C., Cagnazzo, M., Valenzise, G., Tartaglione, E., & Bae, S.-H. (2025). Compression in 3D Gaussian Splatting: A survey of methods, trends, and future directions. https://doi.org/10.48550/arXiv.2502.19457

Balestrieri, M., Valmori, I., & Montuori, M. (2024). UAS and TLS 3D data fusion for built cultural heritage assessment and the application for St. Catherine Monastery in Ferrara, Italy. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-4-2024, 45-52. https://doi.org/10.5194/isprs-archives-XLVIII-M-4-2024-9-2024

Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (ToG)*, *28*(3), 24. https://doi.org/10.1145/1531326.1531330

Beraldin, J.-A. (2004). Integration of laser scanning and close-range photogrammetry - The last decade and beyond. *Proceedings of the XXth ISPRS Congress*, *35*(5), 428-435.

Bieńkowski, R., & Rutkowski, K. (2022). The use of octree in point cloud analysis with application to cultural heritage. https://doi.org/10.48550/arXiv.2301.06936

Bolognesi, M., Furini, A., Russo, V., Pellegrinelli, A., & Russo, P. (2014). Accuracy of cultural heritage 3D models by RPAS and terrestrial photogrammetry. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5, 113-119. https://doi.org/10.5194/isprsarchives-XL-5-113-2014

Chen, C., Guo, J., Wu, H., Li, Y., & Shi, B. (2021). Performance comparison of filtering algorithms for high-density airborne LiDAR point clouds over complex landscapes. *Remote Sensing*, *13*(14), 2663. https://doi.org/10.3390/rs13142663

Clini, P., Nespeca, R., Angeloni, R., & Coppetta, L. (2024). 3D representation of architectural heritage: A comparative analysis of NeRF, Gaussian splatting, and SfM-MVS reconstructions using low-cost sensors. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W8-2024, 91-98. https://doi.org/10.5194/isprs-archives-XLVIII-2-W8-2024-93-2024

Colomina, I., & Molina, P. (2014). Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92, 79-97. https://doi.org/10.1016/j.isprsjprs.2014.02.013

Dalal, A., Hagen, D., Robbersmyr, K. G., & Knausgård, K. M. (2024). Gaussian splatting: 3D reconstruction and novel view synthesis: A review. *IEEE Access*, 12, 96791-96825. https://doi.org/10.1109/ACCESS.2024.3408318

Demantke, J., Mallet, C., David, N., & Vallet, B. (2011). Dimensionality based scale selection in 3D LiDAR point clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-5/W12, 97-102. https://doi.org/10.5194/isprsarchives-XXXVIII-5-W12-97-2011

Dore, C., & Murphy, M. (2012). Integration of Historic Building Information Modeling (HBIM) and 3D GIS for recording and managing cultural heritage sites. *18th International Conference on Virtual Systems and Multimedia (VSMM)*, 369-376. https://doi.org/10.21427/e7sy-rt81

Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., & Felsberg, M. (2024). RoMa: Robust dense feature matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2523-2533. https://doi.org/10.48550/arXiv.2305.15404

Gagliolo, S., Ausonio, E., Federici, B., Ferrando, I., Passoni, D., & Sguerso, D. (2018). 3D cultural heritage documentation: A comparison between different photogrammetric software and their products. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2, 347-354. https://doi.org/10.5194/isprs-archives-XLII-2-347-2018

Galliani, S., Lasinger, K., & Schindler, K. (2015). Massively parallel multiview stereopsis by surface normal diffusion. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 873-881. https://doi.org/10.1109/ICCV.2015.106

Gaong, G. E. A., Idris, A. N., Luh, L. C., Rahman, A. A. A., Sabri, W. M. S. W. M., & Jalil, A. H. A. (2025). Comparative evaluation of 3D building model using UAV photogrammetry and terrestrial laser scanner (TLS). *Built Environment Journal*, *22*(1), 86-104. https://doi.org/10.24191/bej.v22i1.1066

Guédon, A., & Lepetit, V. (2023). SuGaR: Surface-aligned Gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. https://doi.org/10.48550/arXiv.2311.12775

Günen, M. A., Kulakoğlu, F., & Besdok, E. (2024). Accuracy assessment of UAV-based documentation of archaeological site: Kültepe-Kaneš. *Digital Applications in Archaeology and Cultural Heritage*, 35, e00380. https://doi.org/10.1016/j.daach.2024.e00380

Ham, Y., Michalkiewicz, M., & Balakrishnan, G. (2024). DRAGON: Drone and Ground Gaussian Splatting for 3D building reconstruction. *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*, 1-12. https://doi.org/10.48550/arXiv.2407.01761

Hartley, R., & Zisserman, A. (2004). Multiple view geometry in computer vision (2nd ed.). Cambridge University Press.

Jiang, S., Jiang, C., & Jiang, W. (2020). Efficient structure from motion for large-scale UAV images: A review and a comparison of SfM tools. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 230-251. https://doi.org/10.1016/j.isprsjprs.2020.04.016

Jo, I., Lee, Y., Ham, N., Kim, J., & Kim, J.-J. (2025). A quantitative evaluation of UAV flight parameters for SfM-based 3D reconstruction of buildings. *Applied Sciences*, *15*(13), 7196. https://doi.org/10.3390/app15137196

Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Rota Bulò, S., Richardt, C., Ramanan, D., Scherer, S., & Kontschieder, P. (2025). MapAnything: Universal feed-forward metric 3D reconstruction. https://doi.org/10.48550/arXiv.2509.13414

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, *42*(4), 1-14. https://doi.org/10.1145/3592433

Kersten, T., Acevedo Pardo, C., & Lindstaedt, M. (2004). 3D acquisition, modelling and visualization of North German castles by digital architectural photogrammetry. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXV(B2), 126-132.

Kingsland, K. (2020). Comparative analysis of digital photogrammetry software for cultural heritage. *Digital Applications in Archaeology and Cultural Heritage*, 18, e00157. https://doi.org/10.1016/j.daach.2020.e00157

Lague, D., Brodu, N., & Leroux, J. (2013). Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (N-

Z). *ISPRS Journal of Photogrammetry and Remote Sensing*, 82, 10-26. https://doi.org/10.1016/j.isprsjprs.2013.04.009

Li, Q., Yang, G., Gao, C., Huang, Y., Zhang, J., Huang, D., Zhao, B., Chen, X., & Chen, B. M. (2024). Single drone-based 3D reconstruction approach to improve public engagement in conservation of heritage buildings: A case of Hakka Tulou. *Journal of Building Engineering*, 87, 108954. https://doi.org/10.1016/j.jobe.2024.108954

Li, Z., Müller, T., Evans, A., Taylor, R. H., Unberath, M., Liu, M.-Y., & Lin, C.-H. (2023). Neuralangelo: High-fidelity neural surface reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12256-12265. https://doi.org/10.48550/arXiv.2306.03092

Luo, J., Huang, T., Wang, W., & Feng, W. (2024). A review of recent advances in 3D Gaussian splatting for optimization and reconstruction. *Image and Vision Computing*, 151, 105304. https://doi.org/10.1016/j.imavis.2024.105304

Martin-Brualla, R., Radwan, N., Sajjadi, M. S., Barron, J. T., Dosovitskiy, A., & Duckworth, D. (2021). NeRF in the wild: Neural radiance fields for unconstrained photo collections. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7210-7219. https://doi.org/10.48550/arXiv.2008.02268

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, *65*(1), 99-106. https://doi.org/10.1145/3503250

Nex, F., & Remondino, F. (2014). UAV for 3D mapping applications: A review. *Applied Geomatics*, *6*(1), 1-15. https://doi.org/10.1007/s12518-013-0120-x

Owda, A., Balsa-Barreiro, J., & Fritsch, D. (2018). Methodology for digital preservation of the cultural and patrimonial heritage: Generation of a 3D model of the Church St. Peter and Paul (Calw, Germany) by using laser scanning and digital photogrammetry. *Sensor Review*, *38*(3), 282-288. https://doi.org/10.1108/SR-06-2017-0106

Pan, L., Baráth, D., Pollefeys, M., & Schönberger, J. L. (2024). Global structure-from-motion revisited. https://doi.org/10.1007/978-3-031-73661-2_4

Pattee, A., Seitz, C., & Höfle, B. (2015). Integrative 3D recording methods of historic architecture: Burg Hohenecken castle from southwest Germany. *Proceedings of the 2015 Digital Heritage International Congress*, 341-343. https://doi.org/10.1109/DigitalHeritage.2015.7413843

Peng, H., Li, H., Dai, Y., Lan, Y., Luo, Y., Qi, T., Zhang, Z., Zhan, Y., Zhang, J., Xu, W., & Liu, Z. (2025). OmniVGGT: Omni-modality driven visual geometry grounded transformer. https://doi.org/10.48550/arXiv.2511.10560

Pepe, M., Alfio, V. S., & Costantino, D. (2022). UAV platforms and the SfM-MVS approach in the 3D surveys and modelling: A review in the cultural heritage field. *Applied Sciences*, *12*(24), 12886. https://doi.org/10.3390/app122412886

Pritchard, D., Griffo, M., Attenni, M., Barni, R., Bianchini, C., Inglese, C., & Ley, J. (2023). Evolution of recording methods: The Aachen Cathedral World Heritage site documentation project. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-2-2023, 1241-1248. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-1241-2023

Pritchard, D. K., Sperner, J., Hoepner, S., & Tenschert, R. (2017). Terrestrial laser scanning for heritage conservation: The Cologne Cathedral documentation project. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W2, 213-220. https://doi.org/10.5194/isprs-annals-IV-2-W2-213-2017

Remondino, F., Barazzetti, L., Nex, F., Scaioni, M., & Sarazzi, D. (2011). UAV photogrammetry for mapping and 3D modeling – Current status and future perspectives. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-1/C22, 25-31. https://doi.org/10.5194/isprsarchives-XXXVIII-1-C22-25-2011

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). SuperGlue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4938-4947. https://doi.org/10.48550/arXiv.1911.11763

Schönberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104-4113. https://doi.org/10.1109/CVPR.2016.445

Schönberger, J. L., Zheng, E., Pollefeys, M., & Frahm, J.-M. (2016). Pixelwise view selection for unstructured multi-view stereo. *Proceedings of the European Conference on Computer Vision (ECCV)*, 501-518. https://doi.org/10.1007/978-3-319-46484-8_31

Stathopoulou, E. K., & Remondino, F. (2023). A survey on conventional and learning-based methods for multi-view stereo. *The Photogrammetric Record*, *38*(184), 374-407. https://doi.org/10.1111/phor.12456

Stuart, L. A. G., & Pound, M. P. (2025). 3DGS-to-PC: Convert a 3D Gaussian Splatting scene into a dense point cloud or mesh. https://doi.org/10.48550/arXiv.2501.07478

Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., Wang, T., Kristofferson, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., & Kanazawa, A. (2023). Nerfstudio: A modular framework for neural radiance field development. *ACM SIGGRAPH 2023 Conference Proceedings*. https://doi.org/10.1145/3588432.3591516

UNESCO/PERSIST Content Task Force. (2016). The UNESCO/PERSIST guidelines for the selection of digital heritage for long-term preservation. UNESCO.

Wang, J., Chen, M., Karaev, N., Rupprecht, C., Vedaldi, A., & Novotny, D. (2025). VGGT: Visual geometry grounded transformer. https://doi.org/10.48550/arXiv.2503.11651

Wang, J., Karaev, N., Rupprecht, C., & Novotny, D. (2024). VGGSfM: Visual geometry grounded deep structure from motion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21686-21697. https://doi.org/10.48550/arXiv.2312.04563

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., & Revaud, J. (2024). DUSt3R: Geometric 3D vision made easy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20697-20709. https://doi.org/10.48550/arXiv.2312.14132

Weinmann, M. (2016). Reconstruction and analysis of 3D scenes: From irregularly distributed 3D points to object classes. *Springer*.

https://doi.org/10.1007/978-3-319-29246-5

Yu, Y., Verbree, E., van Oosterom, P., Pottgiesser, U., Peng, Y., & Poux, F. (2025). From comparison to integration: A workflow evaluation of 3D Gaussian splatting and LiDAR point cloud for modern architectural heritage. *Automation in Construction*, *180*, 106509. https://doi.org/10.1016/j.autcon.2025.106509

Zachos, A., & Anagnostopoulos, C.-N. (2023). Using terrestrial laser scanning, unmanned aerial vehicles and mixed reality methodologies for digital survey, 3D modelling and historical recreation of religious heritage monuments. *Heritage*, *6*(2), 1642-1665. https://doi.org/10.48550/arXiv.2401.01380

Zhou, Q.-Y., Park, J., & Koltun, V. (2018). Open3D: A modern library for 3D data processing. https://doi.org/10.48550/arXiv.1801.09847

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1851-1858. https://doi.org/10.48550/arXiv.1704.07813