

**MINISTÉRIO DA DEFESA  
EXÉRCITO BRASILEIRO  
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA  
INSTITUTO MILITAR DE ENGENHARIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE DEFESA**

**THIAGO JOÃO  
MIRANDA BALDIVIESO**

**HYBRID MULTIMODAL 3D RECONSTRUCTION FROM UAV  
RGB-THERMAL-LIDAR DATA WITH UNSUPERVISED DEEP EMBEDDINGS**

**RIO DE JANEIRO  
2025**

THIAGO JOÃO  
MIRANDA BALDIVIESO

HYBRID MULTIMODAL 3D RECONSTRUCTION FROM UAV  
RGB-THERMAL-LIDAR DATA WITH UNSUPERVISED DEEP  
EMBEDDINGS

Tese apresentada ao Programa de Pós-graduação em Engenharia de Defesa do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Doutor em Ciências em Engenharia de Defesa.

Orientador(es): Paulo Fernando Ferreira Rosa, Ph.D.  
Luiz Carlos Pacheco Rodrigues Velho,  
Ph.D.

Rio de Janeiro  
2025

©2025

INSTITUTO MILITAR DE ENGENHARIA

Praça General Tibúrcio, 80 – Praia Vermelha

Rio de Janeiro – RJ CEP: 22290-270

Este exemplar é de propriedade do Instituto Militar de Engenharia, que poderá incluí-lo em base de dados, armazenar em computador, microfilmar ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es) e do(s) orientador(es).

João, Thiago; Baldivieso, Miranda.

HYBRID MULTIMODAL 3D RECONSTRUCTION FROM UAV RGB-THERMAL-LIDAR DATA WITH UNSUPERVISED DEEP EMBEDDINGS /

Thiago João e Miranda Baldivieso. – Rio de Janeiro, 2025.

161 f.

Orientador(es): Paulo Fernando Ferreira Rosa e Luiz Carlos Pacheco Rodrigues Velho.

Tese (doutorado) – Instituto Militar de Engenharia, Engenharia de Defesa, 2025.

1. Reconstrução 3D; Sistemas de Aeronaves Remotamente Pilotadas (SARP);  
Aprendizado de Máquina; Integração de Dados Multimodais. i. Ferreira Rosa,  
Paulo Fernando (orient.) ii. Pacheco Rodrigues Velho, Luiz Carlos (orient.) iii.  
Título

**THIAGO JOÃO  
MIRANDA BALDIVIESO**

**HYBRID MULTIMODAL 3D RECONSTRUCTION  
FROM UAV RGB-THERMAL-LIDAR DATA WITH  
UNSUPERVISED DEEP EMBEDDINGS**

Tese apresentada ao Programa de Pós-graduação em Engenharia de Defesa do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Doutor em Ciências em Engenharia de Defesa.

Orientador(es): Paulo Fernando Ferreira Rosa e Luiz Carlos Pacheco Rodrigues Velho.

Aprovada em 24 de Novembro de 2025, pela seguinte banca examinadora:

---

Prof. **PAULO FERNANDO FERREIRA ROSA**  
- Ph.D. do IME - Presidente

---

Prof. **LUIZ CARLOS PACHECO RODRIGUES VELHO**  
- Ph.D. do IMPA

---

Prof. **MAX SUELL DUTRA** - Dr.-Ing. da UFRJ

---

Prof. **MILENA FARIA PINTO** - Ph.D. da CEFET

---

Prof. **RONALDO RIBEIRO GOLDSCHMIDT** - D.Sc. do IME

---

Prof. **DANIEL RODRIGUES DOS SANTOS** - D.Sc. do IME.

Rio de Janeiro  
2025

*I dedicate this work to all who believe in the power of ideas  
and in building something greater through knowledge.*

# ACKNOWLEDGEMENTS

I thank God for everything He has done throughout my journey—for the health, care, wisdom, protection, and motivation to keep studying.

To my family, for their understanding and support during this period of intense dedication.

To my advisors, Paulo Rosa and Luiz Velho, for their availability, attention, guidance, and patience.

I'm grateful to my fellow graduate program colleagues—Fábio Suim, Daniel Morais, Bruno Eduardo, Fabio Luiz, Leandro Moreira, Erick Menezes, Ricardo Maróquio, and Luiz Silva—for their companionship during long hours in the lab, their suggestions for improvement, and their support during experimental missions.

Special thanks go to the members of the Laboratory of Artificial Intelligence, Robotics, and Cybernetics (LIARC), as well as to the professors and staff of the Computer Engineering Section (SE-10) at the Military Institute of Engineering (IME), and to the team at VISGRAF, the computer vision lab at the Institute of Pure and Applied Mathematics (IMPA).

To the professors who kindly accepted the invitation to join the evaluation committee for this work.

It's also important to note that this research was supported by computational infrastructure funded through the Cybernetics Research Subproject, part of the Brazilian Army's Strategic Project.

Finally, I would like to acknowledge the financial support provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) through the Programa de Cooperação Acadêmica – Defesa (PROCAD-DEFESA), Call No. 05/2019, under Project No. 88887.387828/2019-00, entitled “Desenvolvimento de um Sistema de Aeronaves Remotamente Pilotadas com Controle baseado em Alocação Dinâmica para Cobertura de Áreas com Prioridades de Interesse.” I also recognize the doctoral scholarship granted by CAPES within the scope of PROCAD-DEFESA (Process No. 88887.518730/2020-00), linked to the Graduate Program in Defense Engineering (PPG-31007015011P8), which supported the development of this research from October 2020 to September 2024.

Furthermore, I am grateful to the project “Enxame de veículos autônomos aéreos e terrestres: guiamento, controle e navegação (EVAAT-GCN),” initiated in 2025 and registered under Project No. 3311/24, supported by the Fundação de Apoio ao Desenvolvimento da Computação Científica (FADCC). This project is dedicated to the development of

an integrated framework for the coordinated guidance, control, and navigation of heterogeneous autonomous robots operating in both aerial and terrestrial environments with varying levels of automation. Led by a multidisciplinary team of researchers in Artificial Intelligence and Robotics from IME, UFPE, LNCC, UFAM, and IMPA, the EVAAT-GCN initiative has provided collaboration and resources that contributed to the progress of this research.

*"Imagination is more important than knowledge.  
Knowledge is limited. Imagination encircles the world."  
Albert Einstein.*

# RESUMO

O uso crescente de veículos aéreos não tripulados (VANTs), equipados com sensores RGB, térmicos e LiDAR, tem ampliado aplicações em inspeção estrutural, levantamento topográfico e monitoramento ambiental. Reconstruções tridimensionais obtidas a partir de uma única modalidade, contudo, ainda apresentam limitações em precisão e fidelidade estrutural.

Este trabalho propõe a arquitetura híbrida **GaussianFusion**, que realiza integração multimodal precoce, agrupamento estrutural não supervisionado e renderização contínua via *Gaussian-based splatting*. Embora o nome traga “Fusion”, a proposta adota uma lógica de integração modular, permitindo que informações visuais, térmicas e de profundidade interajam desde o início de forma leve e extensível. Essa modularidade abre espaço para novos sensores e aplicações em sistemas robóticos.

A validação envolveu três cenários urbanos reais: CTEEx, Outeiro da Glória e Quinta da Boa Vista. Comparações entre nuvens de pontos de diferentes combinações sensoriais foram realizadas utilizando métricas geométricas (RMSE, Chamfer, Hausdorff) e índices de agrupamento (Silhouette, Davies-Bouldin). Integrações RGB-LiDAR apresentaram os menores desvios geométricos; dados térmicos aumentaram a interpretabilidade sem comprometer a acurácia. Os agrupamentos obtidos foram estruturais, refletindo propriedades como elevação, densidade e emissividade.

No conjunto Glória, a metodologia de *Gaussian-based splatting* foi avaliada para renderização contínua da fachada da igreja. Os testes demonstraram potencial de visualização fotorealística e continuidade estrutural, mas também revelaram limitações quando a densidade angular das vistas era insuficiente, indicando sensibilidade do método à distribuição de pontos de observação.

Foi também realizado um benchmark com o modelo Point-BERT, comparando projeções latentes, métricas de agrupamento e custos computacionais. Os resultados demonstram que o sistema híbrido alcança desempenho competitivo em agrupamento não supervisionado, equilibrando qualidade estrutural e eficiência operacional.

A principal contribuição científica está em demonstrar que a integração multimodal precoce, aliada ao agrupamento estrutural não supervisionado e à renderização contínua, fornece consistência geométrica e radiométrica em ambientes complexos. A arquitetura apresenta aplicabilidade em engenharia civil, agricultura de precisão, avaliação de impactos climáticos e segurança e defesa, mostrando consistência, escalabilidade e capacidade de generalização para modelagem tridimensional em ambientes reais.

**Palavras-chave:** Reconstrução 3D; UAV; Aprendizado de Máquina; Integração Multimodal de Dados.

# ABSTRACT

The growing use of unmanned aerial vehicles (UAVs) equipped with RGB optical sensors, thermal cameras, and LiDAR has expanded applications in structural inspection, topographic surveying, and environmental monitoring. Reconstructions based on a single modality, however, still face limitations in accuracy and structural fidelity.

This thesis proposes the hybrid architecture **GaussianFusion**, which performs early multimodal integration, unsupervised structural clustering, and continuous rendering via *Gaussian-based splatting*. Although the name includes “Fusion,” the proposal adopts a modular integration approach, enabling visual, thermal, and depth information to interact from the outset in a lightweight and extensible manner. This modularity facilitates the incorporation of new sensors and applications in robotic systems.

Validation was carried out in three real urban scenarios: CTE<sub>x</sub>, Glória, and Quinta da Boa Vista. Comparisons among point clouds from different sensor combinations were performed using geometric metrics (RMSE, Chamfer, Hausdorff) and clustering indices (Silhouette, Davies-Bouldin). RGB-LiDAR integrations achieved the lowest geometric deviations; thermal data increased interpretability without compromising accuracy. The resulting clusters were structural, reflecting properties such as elevation, density, and emissivity.

On the Glória dataset, *Gaussian-based splatting* was evaluated for continuous rendering of the church façade. The tests demonstrated photorealistic visualization and structural continuity, but also revealed limitations when angular density was insufficient, highlighting the method’s sensitivity to viewpoint distribution.

A benchmark study was also conducted with the Point-BERT model, comparing latent projections, clustering metrics, and computational costs. Results show that the hybrid system achieves competitive performance in unsupervised clustering, balancing structural quality and operational efficiency.

The main scientific contribution lies in demonstrating that early multimodal integration, combined with unsupervised structural clustering and continuous rendering, provides geometric and radiometric consistency in complex environments. The proposed architecture is applicable to civil engineering, precision agriculture, climate impact assessment, and security and defense, showing consistency, scalability, and generalization capability for 3D modeling in real environments.

**Keywords:** 3D Reconstruction; UAV; Machine Learning; Multimodal Data Integration.

# LIST OF FIGURES

Figure 1 – Conceptual illustration of multimodal UAV acquisition and processing. The colored cone represents the sensor coverage integrating RGB, thermal and LiDAR channels, which can be combined in a integration and continuous rendering pipeline. Illustrative parameters: typical flight altitude 40–80 m; field of view and cone shape depend on payload and mission profile. . . . .	20
Figure 2 – Evolution of drone flight requests in Brazil between 2020–2024. Source: Adapted from (1). . . . .	28
Figure 3 – SfM–MVS reconstruction pipeline with images. . . . .	32
Figure 4 – Pinhole projection model: mapping of a world point $\mathbf{X}_i$ to its image measurement $\tilde{\mathbf{u}}_{ik}$ . . . . .	33
Figure 5 – Triangulation principle: a 3D point $\mathbf{X}_i$ is recovered by intersecting rays back-projected from multiple calibrated camera views. . . . .	34
Figure 6 – Geometric alignment example illustrating ICP-based registration between source and target point clouds, showing initial pose, correspondence indicators and final fit for diagnostic inspection. . . . .	39
Figure 7 – Operational flow of the continuous representation module using Gaussian Splatting from SfM reconstructions. Source: (2). . . . .	43
Figure 8 – Geometric interpretation of an anisotropic Gaussian splat in 3D space. The mean $\mu_i$ defines the center of the distribution, while the covariance governs its spread and orientation along the axes. Opacity modulates the visual contribution of the splat, illustrated here by the semi-transparent surface. . . . .	45
Figure 9 – Projection of latent embeddings using t-SNE (left) and UMAP (right). Each point represents a region in the 3D scene with similar geometric and radiometric properties. . . . .	49
Figure 10 – Conceptual diagrams of clustering evaluation metrics. Left: Silhouette Coefficient, measuring the relative proximity of a point to its own cluster versus neighboring clusters. Right: Davies–Bouldin Index, quantifying intra-cluster dispersion and inter-cluster separation. . . . .	50
Figure 11 – Simplified conceptual flow of the unsupervised segmentation pipeline. Starting from a 3D point cloud, features are extracted using neural architectures, embeddings are computed, clustering is performed in latent space, and labels are projected back onto the segmented cloud. . . . .	51

Figure 12 – Conceptual pipeline of the <code>GaussianFusion_AI</code> architecture. Each module processes multimodal inputs toward an unsupervised feature-clustered and continuously rendered 3D scene using Gaussian-based splatting. . . . .	69
Figure 13 – Gaussian-based splatting module used in <code>GaussianFusion_AI</code> . The module initializes splats from the integrated cloud, applies image-based texturing, and composes a viewpoint-dependent rendering using soft blending. . . . .	74
Figure 14 – Pipeline for unsupervised feature clustering based on latent embeddings; outputs are clustered point clouds used for downstream analysis and continuous rendering with Gaussian-based splatting. . . . .	75
Figure 15 – Illustration of geometric evaluation metrics. Left: RMSE measures global deviation. Center: Chamfer Distance captures average proximity. Right: Hausdorff Distance highlights worst-case mismatch. . . . .	79
Figure 16 – DJI Matrice 350 RTK and Mavic 3T Enterprise, used in outdoor experiments. Source: (3). . . . .	84
Figure 17 – DJI Pilot 2 interface used for automated multisensor mission planning and route configuration for the CTE <sub>x</sub> surveys. Adapted from manufacturer documentation (4) . . . . .	84
Figure 18 – DJI Mavic 3T Enterprise — annotated external components: propellers and motors, collision sensors, landing gear, battery compartments and RTK antenna. Source: Adapted from (3) . . . . .	85
Figure 19 – DJI Matrice 350 RTK — annotated external components: propellers and motors, collision sensors, landing gear, battery compartments and RTK antennas. Source: Adapted from (3) . . . . .	85
Figure 20 – Zenmuse L2, hybrid sensor used for integrated RGB and LiDAR acquisition. Source: (3) . . . . .	86
Figure 21 – RTK base station with GNSS antenna, radio modem and power module; used to provide differential corrections to airborne RTK receivers. Source: From (3) . . . . .	87
Figure 22 – Per-sensor reconstructions and pairwise integrations for the CTE <sub>x</sub> dataset, focused on a goalpost and a heat-emitting box. (a) RGB-only cloud. (b) RGB + thermal integration. (c) RGB + LiDAR integration. (d) Thermal-only cloud. (e) Thermal + LiDAR integration. (f) LiDAR-only cloud. All views depict the same cropped region to compare the two scene elements across sensing modalities. . . . .	89
Figure 23 – RGB–Thermal–LiDAR integrated cloud for CTE <sub>x</sub> (lateral crop). . . . .	90
Figure 24 – Feature Clustering with PointNet + k-means. Artificial clusters and vertical distortions over flat areas. . . . .	92

Figure 25 – Latent projections for CTE <sub>x</sub> : UMAP (up) and t-SNE (down) coloured by attributes. Continuous core, no clearly separable clusters. . . . .	93
Figure 26 – Gloria reconstructions: (a,b) RGB; (c,d) Thermal; (e,f) RGB–Thermal integration (zoom on the church). . . . .	95
Figure 27 – 3D UMAP projection of embeddings — Gloria dataset. . . . .	97
Figure 28 – t-SNE projection of embeddings — Gloria dataset. . . . .	97
Figure 29 – Continuous rendering of the church side wall using Gaussian-based splatting. The offset expands coverage but also reveals sensitivity to viewpoint distribution, highlighting the contribution of the method in achieving photorealistic visualization while exposing its dependence on angular density. . . . .	99
Figure 30 – Continuous rendering with elevated lateral offset, showing structural coherence loss in the Gaussian representation. This illustrates the contribution of the experiment in identifying the operational limits of Gaussian-based splatting under extreme pose extrapolation. . . . .	100
Figure 31 – Evolution of visual representation in the Gloria dataset, from SfM façade projection to simulated 2D segmentation, continuous rendering with Gaussian-based splatting, and 3D feature clustering of the RGB-Thermal cloud. The sequence demonstrates the contribution of GaussianFusion_AI in transitioning from discrete image segmentation to continuous multimodal inference, achieving structural coherence and enriched visualization. . . . .	101
Figure 32 – Per-sensor reconstructions for Quinta da Boa Vista, focusing on the National Museum façade and roof. The comparison contributes by showing how RGB ensures geometric richness, thermal adds radiometric context, and LiDAR reinforces structural fidelity, while integration enhances multimodal consistency. . . . .	103
Figure 33 – Integration variants for Quinta da Boa Vista, including RGB-Thermal, RGB-LiDAR, Thermal-LiDAR, and RGB-Thermal-LiDAR (zoomed). The contribution lies in illustrating how different sensor combinations affect structural fidelity and radiometric enrichment, confirming that RGB-based integrations yield the most complete models, while triple integration achieves balanced geometric and radiometric consistency. . . . .	104
Figure 34 – Exploratory 3D UMAP projections for QUINTA, highlighting altitude, spectral channels and mean intensity. . . . .	105

Figure 35 – Results of KDTree combined with k-means on the Quinta dataset, showing UMAP 3D embeddings, clustering mapped to the original cloud, and the RGB-Thermal-LiDAR input cloud. The contribution of this experiment is to demonstrate that unsupervised clustering yielded continuous structural partitions related to elevation, density, and emissivity, reinforcing their interpretation as structural rather than semantic groupings. . . . .	107
Figure 36 – UMAP 2D projection, dense core with peripheral latent islands. . . . .	108
Figure 37 – UMAP 3D projection, structural continuity and absence of sharp latent cluster separations. . . . .	108
Figure 38 – t-SNE projection, compressed latent structure with thin, ill-defined branches. . . . .	109
Figure 39 – Overview of the USGS Darby dataset, integrating LiDAR cloud, RGB bands and thermal layer. Source: From (5) . . . . .	112
Figure 40 – UMAP 3D projections of hybrid architecture embeddings, colored by attributes: neutral, elevation (Z), R, G, B, and temperature. . . . .	117
Figure 41 – t-SNE 3D projections of hybrid architecture embeddings, colored by attributes: neutral, elevation (Z), R, G, B, and temperature. . . . .	117
Figure 42 – UMAP 3D projections of PointBERT embeddings, colored by attributes: neutral, elevation (Z), R, G, B, and temperature. . . . .	118
Figure 43 – t-SNE 3D projections of PointBERT embeddings, colored by attributes: neutral, elevation (Z), R, G, B, and temperature. . . . .	118
Figure 44 – UMAP (left) and t-SNE (right) projections of latent encodings for CTE <sub>x</sub> , Gloria and Quinta da Boa Vista (top to bottom). Colors indicate height (Z) and mean RGB intensity. . . . .	125
Figure 45 – Processing pipeline applied to the USGS Darby dataset. . . . .	149
Figure 46 – UMAP projection with HDBSCAN labels for Hybrid, seed 42. . . . .	151
Figure 47 – UMAP projection with HDBSCAN labels for Point-BERT, seed 42. . . . .	152
Figure 48 – UMAP projection with HDBSCAN labels for Hybrid, seed 123. . . . .	152
Figure 49 – UMAP projection with HDBSCAN labels for Point-BERT, seed 123. . . . .	153
Figure 50 – UMAP projection with HDBSCAN labels for Hybrid, seed 2025. . . . .	153
Figure 51 – UMAP projection with HDBSCAN labels for Point-BERT, seed 2025. . . . .	154
Figure 52 – Slide-style flowchart of the Gaussian-based splatting module used in experiments. . . . .	158
Figure 53 – Schematic of covariance projection from 3D to image plane. The implementation uses an isotropic image-space approximation $\sigma_{p,i}$ derived from $\Sigma_i$ ; anisotropic image-plane covariances are noted in the text as an alternative. . . . .	160

## LIST OF TABLES

Table 1 – Comparison between recent works and this proposal . . . . .	64
Table 2 – Technical comparison of the drones used . . . . .	87
Table 3 – Sensors and corresponding acquisition areas . . . . .	87
Table 4 – Pairwise and integrated cloud comparisons — CTE <sub>x</sub> dataset . . . . .	90
Table 5 – Compound integration comparisons — CTE <sub>x</sub> (Batch 5) . . . . .	90
Table 6 – Unsupervised feature clustering performance — CTE <sub>x</sub> dataset (k-means results) . . . . .	91
Table 7 – Comparisons for GL-1 (Gloria): RGB, Thermal and Integrations . . . . .	95
Table 8 – Unsupervised feature clustering performance — Gloria . . . . .	96
Table 9 – Parameters used in lateral rendering . . . . .	99
Table 10 – Structural comparisons between isolated and integrated clouds — QUINTA	105
Table 11 – Compound integration comparisons — QUINTA (Batch 5) . . . . .	106
Table 12 – Feature clustering metrics by architecture — Quinta da Boa Vista . . . . .	106
Table 13 – Clustering comparison (hybrid architecture) . . . . .	114
Table 14 – Processing times per stage (hybrid architecture) . . . . .	115
Table 15 – Clustering comparison with PointBERT . . . . .	115
Table 16 – Processing times with PointBERT . . . . .	115
Table 17 – Consistency across multiple seeds (HDBSCAN after t-SNE) . . . . .	116
Table 18 – Comparative summary of integrations by dataset . . . . .	123
Table 19 – Feature clustering comparison by dataset and architecture (Silhouette) . . . . .	124
Table 20 – Consolidated clustering results across seeds and encoders (HDBSCAN after t-SNE). . . . .	150
Table 21 – Gaussian-based splatting parameters used in the Glória experiment . . . . .	161

# LIST OF ABBREVIATIONS AND ACRONYMS

**CMVS** — Clustering Multi-View Stereo

**COLMAP** — COnstrained Local Minimization and Projective reconstruction

**DB** — Índice de Davies-Bouldin

**DBSCAN** — Density-Based Spatial Clustering of Applications with Noise

**DGCNN** — Dynamic Graph Convolutional Neural Network

**DN** — Digital Number

**EXIF** — Exchangeable Image File Format

**FPS** — Farthest Point Sampling

**GNSS** — Global Navigation Satellite System

**ICP** — Iterative Closest Point

**IMU** — Inertial Measurement Unit

**KD-Tree** — K-Dimensional Tree

**LiDAR** — Light Detection and Ranging

**MVE** — Multi-View Environment

**MVS** — Multi-View Stereo

**NTRIP** — Networked Transport of RTCM via Internet Protocol

**NERF** — Neural Radiance Fields

**OA** — Overall Accuracy

**ORB** — Oriented FAST and Rotated BRIEF

**PLY** — Polygon File Format

**PMVS** — Patch-based Multi-View Stereo

**PPK** — Post-Processed Kinematic

**RGB** — Red, Green, Blue

**RMSE** — Root Mean Square Error

**RTK** — Real-Time Kinematic

**SfM** — Structure from Motion

**SIFT** — Scale-Invariant Feature Transform

**SNT** — Spanning Neighborhood Tree

**t-SNE** — T-Distributed Stochastic Neighbor Embedding

**SURF** — Speeded-Up Robust Features

**UMAP** — Uniform Manifold Approximation and Projection

**UAV** — Unmanned Aerial Vehicle

**VGGT** — Visual Geometry Grounded Transformer

# CONTENTS

<b>1</b>	<b>INTRODUCTION</b>	<b>20</b>
1.1	MOTIVATION	21
1.2	PROBLEM CHARACTERIZATION	22
1.3	RESEARCH QUESTION	22
1.4	HYPOTHESIS	23
1.5	OBJECTIVES	23
1.6	JUSTIFICATION	23
1.7	THESIS STRUCTURE	24
<b>2</b>	<b>THEORETICAL FRAMEWORK</b>	<b>26</b>
2.1	UNMANNED AERIAL VEHICLES AND THE ADVANCE OF MULTIMODAL SENSING	27
2.2	THREE-DIMENSIONAL RECONSTRUCTION FROM IMAGES: FROM PHOTOGRAMMETRY TO COMPUTATIONAL GEOMETRY	30
2.2.1	DIGITAL PHOTOGRAMMETRY AND STRUCTURE-FROM-MOTION (SFM)	31
2.2.1.1	PROJECTION MODEL AND NOTATION	32
2.2.1.2	TRIANGULATION AND SPARSE RECONSTRUCTION	33
2.2.1.3	BUNDLE ADJUSTMENT AND GLOBAL REFINEMENT	33
2.2.2	MULTI-VIEW STEREO: DENSIFYING STRUCTURE	34
2.3	MULTIMODAL INTEGRATION OF THREE-DIMENSIONAL DATA: UNITING GEOMETRY, COLOR AND TEMPERATURE	37
2.3.1	PREPROCESSING, CALIBRATION AND SHARED REFERENCE FRAMES	37
2.3.2	GEOMETRIC CO-REGISTRATION: ALGORITHMS, CONSISTENCY AND DIAGNOSTICS	38
2.3.3	ATTRIBUTE TRANSFER: PROJECTION, GAUSSIAN INTERPOLATION AND HYBRID STRATEGY	39
2.3.4	SERIALIZATION, ENGINEERING AND VALIDATION	42
2.4	CONTINUOUS RENDERING WITH GAUSSIAN SPLATTING	43
2.5	UNSUPERVISED FEATURE CLUSTERING IN 3D CLOUDS: LATENT EMBEDDINGS AND STRUCTURAL COHERENCE	47
<b>3</b>	<b>INTEGRATED REVIEW OF SCIENTIFIC LITERATURE</b>	<b>54</b>
3.1	MULTIMODAL DATA INTEGRATION	55
3.2	CONTINUOUS REPRESENTATIONS AND NEURAL SCENE MODELING	56
3.3	UNSUPERVISED FEATURE CLUSTERING AND 3D REPRESENTATION LEARNING	58

3.4	RECENT ADVANCES IN 3D RECONSTRUCTION AND SEMANTIC MODELING . . . . .	59
3.5	GAPS, PRACTICAL CONSTRAINTS AND REPRODUCIBILITY . . . . .	61
3.6	COMPARATIVE SYNTHESIS AND RATIONALE FOR GAUSSIANFUSION_AI . . . . .	62
3.7	CONTEXTUALIZATION AND SELECTED LITERATURE . . . . .	63
<b>4</b>	<b>HYBRID ARCHITECTURE FOR FEATURE CLUSTERING OF UAV MULTIMODAL DATA . . . . .</b>	<b>66</b>
4.1	THEORETICAL FOUNDATIONS AND RESEARCH FRAMING . . . . .	66
4.2	OVERVIEW OF THE EXPERIMENTAL ARCHITECTURE . . . . .	69
4.3	PREPROCESSING AND MULTIMODAL STRUCTURING . . . . .	70
4.3.1	PHOTOGRAMMETRIC RECONSTRUCTION OF RGB AND THERMAL DATA . . . . .	71
4.3.2	GEOMETRIC PROCESSING OF THE LIDAR CLOUD . . . . .	72
4.3.3	MULTIMODAL INTEGRATION AND CROSS INTERPOLATION . . . . .	72
4.3.4	CONTINUOUS MODELING WITH GAUSSIAN-BASED SPLATTING . . . . .	73
4.4	FEATURE EXTRACTION AND CLUSTERING . . . . .	74
4.4.1	OVERALL FLOW OF FEATURE EXTRACTION AND CLUSTERING . . . . .	75
4.4.2	INPUT VECTOR AND PREPROCESSING . . . . .	75
4.4.3	NEURAL NETWORKS FOR EMBEDDING EXTRACTION . . . . .	76
4.4.4	CLUSTERING IN THE LATENT SPACE . . . . .	76
4.4.5	VISUALIZATION AND QUALITATIVE INSPECTION . . . . .	77
4.5	QUANTITATIVE EVALUATION OF RECONSTRUCTION AND CLUSTERING . . . . .	77
4.5.1	RECONSTRUCTION EVALUATION . . . . .	77
4.5.2	CLUSTERING VALIDATION . . . . .	78
4.5.3	COMPUTATIONAL METRICS AND TIMING ANALYSIS . . . . .	79
4.5.4	EXPERIMENTAL AND COMPARATIVE PROCEDURES . . . . .	79
4.6	COMPUTATIONAL ARCHITECTURE AND IMPLEMENTED ALGORITHMS . . . . .	80
4.6.1	INTEGRATION AND ALIGNMENT OF POINT CLOUDS . . . . .	80
4.6.2	INTEGRATION AND ALIGNMENT OF POINT CLOUDS . . . . .	80
4.6.3	RENDERING WITH GAUSSIAN-BASED SPLATTING . . . . .	81
<b>5</b>	<b>MULTISENSORY EXPERIMENTS . . . . .</b>	<b>83</b>
5.1	COMPUTATIONAL INFRASTRUCTURE AND AERIAL PLATFORMS . . . . .	83
5.2	CTEX DATASET — INTEGRATION AND FEATURE CLUSTERING EVALUATION IN A CONTROLLED FIELD . . . . .	88
5.2.1	ACQUISITION AND SENSOR RECONSTRUCTION . . . . .	88
5.2.2	UNSUPERVISED FEATURE CLUSTERING . . . . .	91
5.3	GLORIA DATASET - MULTISENSORY RECONSTRUCTION IN AN URBAN HERITAGE ENVIRONMENT . . . . .	94
5.3.1	RECONSTRUCTION AND INTEGRATION ANALYSIS . . . . .	94

5.3.2	UNSUPERVISED FEATURE CLUSTERING AND LATENT PROJECTIONS . . .	96
5.3.3	CONTINUOUS RENDERING WITH GAUSSIAN-BASED SPLATTING . . . . .	98
5.4	QUINTA DATASET — INTEGRATION AND FEATURE CLUSTERING IN AN URBAN PARK . . . . .	102
5.4.1	ACQUISITION AND SENSOR RECONSTRUCTIONS . . . . .	102
5.4.2	MULTISENSOR INTEGRATION AND QUANTITATIVE ANALYSIS . . . . .	105
<b>6</b>	<b>UNSUPERVISED FEATURE CLUSTERING AND BENCHMARKING</b>	<b>111</b>
6.1	DATASET DESCRIPTION . . . . .	111
6.2	PREPROCESSING AND TOOLS . . . . .	113
6.3	TOKENIZATION AND PARAMETER JUSTIFICATION . . . . .	113
6.4	ARCHITECTURES EVALUATED . . . . .	113
6.5	EXPERIMENTAL SETUP . . . . .	114
6.6	QUANTITATIVE RESULTS . . . . .	114
6.7	SPATIAL INTERPRETATION OF PROJECTIONS . . . . .	116
6.8	ABLATION STUDY . . . . .	119
6.8.1	OPERATIONAL VIABILITY AND BOTTLENECKS . . . . .	120
6.8.2	LIMITATIONS AND FUTURE DIRECTIONS . . . . .	120
<b>7</b>	<b>COMPARATIVE RESULTS AND GENERALIZATION . . . . .</b>	<b>122</b>
7.1	CROSS-DATASET COMPARISON OF MULTIMODAL INTEGRATIONS . .	122
7.2	PERFORMANCE OF UNSUPERVISED FEATURE CLUSTERING STRATE- GIES . . . . .	123
7.3	LATENT MAPS AND COMPARATIVE STRUCTURE OF EMBEDDINGS .	124
7.4	INTEGRATION WITH BENCHMARK FINDINGS . . . . .	126
7.5	FAILURE MODES AND DIAGNOSTIC INTERPRETATION . . . . .	127
7.6	SYNTHESIS ACROSS EXPERIMENTAL CONTEXTS . . . . .	128
<b>8</b>	<b>CONCLUSIONS . . . . .</b>	<b>130</b>
8.1	FUTURE PERSPECTIVES . . . . .	133
	<b>BIBLIOGRAPHY . . . . .</b>	<b>135</b>
	<b>APPENDIX A — ACADEMIC AND TECHNICAL PUBLICATIONS</b>	<b>146</b>
	<b>APPENDIX B — ROBUSTNESS ANALYSIS OF UNSUPERVISED CLUSTERING ON THE USGS DARBY DATASET</b>	<b>148</b>
	<b>APPENDIX C — GAUSSIAN-BASED SPLATTING MODULE: IM- PLEMENTATION DETAILS AND FLOWCHART</b>	<b>157</b>

# 1 INTRODUCTION

Three-dimensional (3D) reconstruction of real-world environments is a central field in computer vision, driven by recent advances in deep learning, remote sensing, and neural rendering (6, 7). This area has grown in importance due to applications in urban mapping, environmental monitoring, and infrastructure inspection, particularly when combined with unmanned aerial vehicles (UAVs) (8).

Figure 1 illustrates the conceptual workflow of multimodal UAV acquisition and processing, highlighting how RGB, thermal, and LiDAR channels can be integrated into a unified pipeline for continuous rendering and multimodal integration. This visual representation helps contextualize the methodological challenges addressed in this work. The illustration was optimized with the aid of image editing software and Google Gemini, based on concepts and specifications developed in this research.



Figure 1 – Conceptual illustration of multimodal UAV acquisition and processing. The colored cone represents the sensor coverage integrating RGB, thermal and LiDAR channels, which can be combined in a integration and continuous rendering pipeline. Illustrative parameters: typical flight altitude 40–80 m; field of view and cone shape depend on payload and mission profile.

Despite progress, challenges remain, such as efficient integration of data from heterogeneous sensors and consistency under adverse environmental conditions (9). These difficulties are particularly relevant in dense and complex urban scenarios, where occlusions, variable lighting, and material diversity impose additional constraints on reconstruction

pipelines. The explicit focus on real urban environments is therefore justified: such contexts represent some of the most demanding and impactful applications, directly linked to smart cities, infrastructure monitoring, and disaster response. By addressing these challenges, solutions developed here can later be generalized to other domains.

Recent techniques such as Gaussian Splatting (2, 10) and architectures including PointNet (11), PointNet++ (12), DGCNN (13), KD-tree and Point-MAE (14), when applied in an unsupervised regime, have shown substantial potential for 3D segmentation, latent representation learning, and continuous model generation from point clouds.

In this context, we propose `GaussianFusion_AI`, a hybrid and modular architecture for 3D reconstruction and segmentation of urban environments based on the integration of RGB, thermal and LiDAR data collected by UAVs. The name reflects the integration of differentiable rendering via Gaussian primitives with artificial intelligence (AI) methods for unsupervised feature clustering. The project was designed to be open, scalable and reproducible; code and experiments are publicly available at [https://gitlab.com/tjmb\\_ime/GaussianFusion\\_AI](https://gitlab.com/tjmb_ime/GaussianFusion_AI), reinforcing the principles of open science and encouraging collaboration among researchers in the field.

## 1.1 Motivation

The primary motivation of this work is the persistent gap between multimodal UAV data acquisition and the creation of realistic, semantically informative 3D models for practical use. Although RGB, thermal, and LiDAR sensors provide complementary signals, effective integration remains hindered by spatial and radiometric misalignment, redundant or inconsistent measurements, and heterogeneous noise profiles that complicate unified representation and downstream analysis (8, 9).

RGB imagery provides high-resolution texture and color cues but suffers from illumination variability and occlusions. Thermal sensors capture emissivity and temperature gradients, which are valuable for anomaly detection and environmental monitoring, yet they typically operate at lower spatial resolution. LiDAR delivers precise geometric structure and metric accuracy, but lacks semantic richness. Integrating these modalities is therefore essential: each compensates for the limitations of the others, enabling reconstructions that are simultaneously geometrically accurate, radiometrically expressive, and semantically interpretable.

There is a growing need for interpretable 3D models that support operational decision-making in domains such as structural inspection, urban planning, and disaster response. Meeting this need requires methods that jointly optimize geometric accuracy, semantic discrimination, and computational efficiency. This research addresses that requirement by investigating sensor integration strategies, deep-learning-based unsupervised

feature clustering, and differentiable continuous rendering as the core pillars of the proposed GaussianFusion\_AI architecture.

## 1.2 Problem Characterization

Despite advances in UAV data acquisition, multimodal 3D reconstruction still faces significant obstacles:

- **Sensor heterogeneity:** Different sensors produce data with varying resolutions, sampling patterns, formats, and noise characteristics, which complicates direct integration and joint processing (8, 9).
- **Spatial alignment:** Precise registration between RGB imagery, thermal imagery, and LiDAR point clouds demands robust extrinsic calibration, multi-sensor correspondence, and compensation for timing and viewpoint differences (8).
- **Geometric representation:** Mesh- and voxel-based approaches often trade fidelity for scalability; continuous representations are desirable but require efficient, multimodal-aware parameterizations (2, 10).
- **Unsupervised representation learning:** Learning useful latent embeddings and clusters from multimodal 3D data must handle unstructured, high-dimensional inputs and cope with modality-specific ambiguities, noise, and scale variation (11, 12, 13, 14).

The absence of integrated pipelines that jointly address sensor integration, unsupervised representation learning, and continuous differentiable rendering constrains real-time deployment and robust operation in complex, dynamic environments.

## 1.3 Research Question

*How can a hybrid architecture that integrates multimodal data (RGB, thermal, LiDAR), unsupervised representation learning, and differentiable continuous rendering improve the accuracy, completeness, and interpretability of 3D reconstructions obtained from UAV platforms in complex urban environments?*

This question is motivated by the limitations of classical fusion approaches, which rely primarily on sensor calibration and geometric registration (8). While such methods reduce local minima in non-linear optimization, they remain sensitive to scene complexity (e.g., repetitive façades, flat roofs, vegetation, reflective surfaces) and often fail to produce semantically rich models. Recent advances in point cloud learning (11, 12, 13, 14) and continuous Gaussian-based rendering (2, 10) highlight opportunities to overcome these gaps, but integrated multimodal UAV pipelines remain scarce.

## 1.4 Hypothesis

*By combining multimodal integration of UAV-acquired RGB, thermal, and LiDAR data with unsupervised deep-learning-based feature clustering and differentiable continuous rendering, it is possible to generate 3D reconstructions that are more geometrically stable, structurally informative, and computationally scalable than those produced by mono-sensor or calibration-only approaches.*

## 1.5 Objectives

### General Objective

Develop a hybrid, scalable architecture for 3D reconstruction and structural clustering of real urban environments by integrating UAV-acquired RGB, thermal, and LiDAR data, combining unsupervised deep-learning encoders with differentiable continuous rendering via Gaussian-based splatting.

### Specific Objectives

- Perform calibration and geometric registration of point clouds derived from RGB, thermal, and LiDAR sensors mounted on UAVs, establishing a shared reference frame.
- Implement and compare unsupervised 3D feature representation and clustering methods using PointNet, PointNet++, DGCNN, Point-MAE, and KD-tree-based approaches on the integrated point clouds.
- Integrate a continuous Gaussian-based splatting representation into the visualization and inference pipeline to model smooth, projectable surfaces.
- Quantitatively evaluate reconstruction and clustering using geometric metrics (RMSE, Chamfer Distance, Hausdorff Distance) and structural indicators (Silhouette score, intra-cluster consistency).
- Release source code, datasets, and experimental scripts in the public repository [https://gitlab.com/tjmb\\_ime/GaussianFusion\\_AI](https://gitlab.com/tjmb_ime/GaussianFusion_AI) to ensure transparency, reproducibility, and alignment with open science practices.

## 1.6 Justification

Multimodal 3D reconstruction using data from unmanned aerial systems is a promising technological frontier with direct applications in engineering, defense, agriculture,

and cultural heritage. Although UAV platforms increasingly carry optical, thermal, and LiDAR sensors, few integrated solutions effectively combine precise sensor calibration, multimodal data integration, unsupervised structural clustering, and continuous rendering for complex urban environments.

This work differentiates itself from the state of the art by articulating unsupervised deep learning with differentiable neural rendering, producing models that capture both geometry and structural attributes. The native integration of multiple sensory domains, together with continuous representations based on Gaussian-based splatting, addresses gaps in conventional methods and aims to deliver reconstructions that are more complete, interpretable, and scalable.

## 1.7 Thesis Structure

The organization of this thesis follows a sequential logic that reflects the methodological trajectory adopted throughout the research. Chapter 1 introduces the motivation, problem characterization, research question, objectives, justification, expected contributions, and the overall structure of the work.

Chapter 2 presents the theoretical framework, covering the evolution of unmanned aerial vehicles and multimodal sensing, the principles of three-dimensional reconstruction from images, and the foundations of multimodal data integration. It also discusses continuous rendering with Gaussian-based splatting and unsupervised feature clustering in 3D point clouds.

Chapter 3 provides an integrated review of the scientific literature, including recent advances in multimodal data integration, continuous representations, neural scene modeling, and unsupervised 3D representation learning. It identifies practical gaps and reproducibility challenges, and synthesizes the rationale behind the proposed `GaussianFusion_AI` architecture.

Chapter 4 details the hybrid methodology developed, emphasizing the stages of acquisition, preprocessing, multimodal integration, continuous modeling with Gaussian-based splatting, and unsupervised feature clustering. It also presents the metrics used for quantitative and structural evaluation.

Chapter 5 describes the multisensory experiments conducted in three urban scenarios—CTEx, Outeiro da Glória, and Quinta da Boa Vista—highlighting the acquisition strategies, integration results, clustering outcomes, and rendering performance in each context.

Chapter 6 presents the experimental evaluation and benchmarking, including dataset descriptions, preprocessing tools, architectural comparisons, quantitative results,

spatial interpretation of latent projections, and an ablation study.

Chapter 7 consolidates comparative results and generalization analysis, comparing multimodal integration strategies across datasets, evaluating clustering performance, and interpreting latent structures. It also discusses failure modes and synthesizes insights across experimental contexts.

Chapter 8 concludes the thesis by summarizing findings, outlining future perspectives, and compiling academic and technical publications derived from the project. Appendices provide additional benchmarking results and documentation of related work.

## Conclusion

Three-dimensional reconstruction of real environments through the integration of optical, thermal, and LiDAR data acquired by UAVs represents a rapidly evolving field, shaped by recent advances and persistent practical challenges. This thesis proposes a hybrid approach that combines unsupervised deep learning, multimodal data integration, and continuous rendering via Gaussian-based splatting, aiming to deliver a coherent and adaptable solution for complex urban contexts.

The work is grounded in experiments with real data, the application of established geometric and structural metrics, and the use of open-source tools, with a strong emphasis on reproducibility and practical relevance. The following chapters present the theoretical foundations, related works, methodological design, and experimental results that support and validate the proposed architecture.

## 2 THEORETICAL FRAMEWORK

This chapter establishes the theoretical and conceptual foundations that underpin the methodology proposed in this thesis. Drawing on recent advances in Remotely Piloted Aerial Systems (RPAS), multimodal sensing, and differentiable rendering, it explores the principles that support three-dimensional reconstruction of real environments, fusion among heterogeneous sensor modalities, and semantic inference in unlabeled point clouds. The goal is to ground methodological choices in the current scientific literature while identifying gaps that justify the development of the `GaussianFusion_AI` architecture.

The chapter is organized into six main sections. The first presents the technological evolution of RPAS and their growing application in high-resolution urban sensing. The second examines three-dimensional modeling techniques based on digital photogrammetry, with emphasis on the SfM–MVS paradigm. The third delves into multimodal data integration strategies among RGB, thermal and LiDAR data, addressing their operational benefits and challenges. The fourth section discusses the concept of continuous representation via Gaussian Splatting, articulating its mathematical foundations and application to urban modeling. The fifth section describes contemporary approaches to unsupervised feature clustering in 3D point clouds, emphasizing self-supervised architectures and vector clustering techniques. Finally, the sixth section provides a critical literature review and highlights methodological gaps that motivated this research.

Multimodal sensing has become a central theme in contemporary photogrammetry and remote sensing, motivated by the limitations of single-modality approaches in complex urban and environmental scenarios. While RGB imagery provides high-resolution texture and color information, it is often insufficient in shadowed or texture-poor regions. Thermal imaging contributes emissivity and contextual cues related to material properties and environmental conditions, whereas LiDAR supplies precise geometric and topographic measurements. The integrated use of these modalities—sometimes referred to in the literature as multimodal or multi-sensor photogrammetry (15, 16)—has been explored in works published in ISPRS, Photogrammetric Record and IEEE Geoscience and Remote Sensing Letters, highlighting its potential for urban monitoring, heritage documentation and environmental diagnostics. In this thesis, the expression *aerofotogrametria multimodal* is adopted to denote the coordinated acquisition and fusion of heterogeneous aerial sensor data, aligning with these research trends and providing the conceptual foundation for the `GaussianFusion_AI` architecture.

**Terminological note:**

Throughout this work, the architecture is named `GaussianFusion_AI`, preserving the original designation adopted during development. The term *fusion*, present in section titles and module references, should be interpreted conceptually as *early multimodal integration*, emphasizing the modular and lightweight character of the system rather than traditional sensor fusion pipelines. The use of the word “fusion” is therefore historical, restricted to the name of the architecture, while all methodological descriptions adopt the notion of integration. This distinction highlights the ability of the architecture to incorporate new modalities and operate in embedded or robotic contexts.

In addition, the term *continuous representation* refers to 3D models that replace discrete structures with smooth spatial functions. In Gaussian Splatting, for example, each point is treated as a differentiable anisotropic Gaussian distribution. The term *continuous rendering* designates the projection of those models into the image domain, producing smooth renderings that are multi-view and compatible with gradient backpropagation. While continuous representation defines the underlying functional model, continuous rendering specifies its visualization in image space.

As highlighted in recent works (2, 10, 13), this terminology reflects the shift from discrete voxel or mesh structures toward differentiable, function-based models. By adopting this conceptual framing, the thesis aligns with contemporary research trends while extending them to multimodal UAV data integration and continuous modeling.

This clarification ensures that all subsequent mentions of the term *fusion* in this thesis, whether in section titles, figures, or methodological descriptions, are to be interpreted within the framework of early multimodal integration.

## 2.1 Unmanned Aerial Vehicles and the Advance of Multimodal Sensing

The dissemination of UAVs has profoundly reshaped remote sensing and geospatial data acquisition practices across multiple scales (17, 18). Initially restricted to military and experimental applications, UAVs have become decisive components in workflows for topography, surveying, architecture, archaeology and civil engineering, enabling high-resolution, flexible and repeatable data capture (8).

This advance is driven by three complementary vectors: first, the miniaturization of onboard sensors that now allows small airframes to carry high-resolution optical cameras, thermal modules and compact LiDAR systems; second, increased flight autonomy and mechanical stability afforded by improved batteries, multirotor designs and advanced positioning solutions such as RTK/PPK GNSS; and third, greater accessibility resulting

from cost reduction, industrial maturation and evolving regulatory frameworks (8, 17, 19). In Brazil, normative instruments and military guidance have helped shape operational practice and safety requirements for routine UAV deployments (19, 20).

Market analyses and sector reports illustrate the rapid expansion of UAV applications in mapping and monitoring, with strong growth projected for domains such as urban infrastructure inspection, precision agriculture, civil security and disaster response areas that increasingly depend on three-dimensional reconstruction and multimodal sensing pipelines (21). These trends reflect both technological readiness and rising institutional demand for actionable geospatial information.

At the national level, the operational uptake of UAVs is evident in administrative statistics. Registered flight requests processed through SARPAS (Airspace Access Request System) increased markedly between 2020 and 2024, signaling a transition from recreational use to predominantly professional and institutional missions focused on mapping, inspection and environmental monitoring (1). Preliminary data from the first quarter of 2025 confirm this upward trend, with a 22% increase in flight requests compared to the same period of 2024, alongside growth in manual and automated analyses of submissions (22). Figure 2 summarizes the growth up to 2024 and highlights the growing operational pressure for standardized multimodal acquisition protocols and reliable processing chains.

## DRONE FLIGHT REQUESTS IN 2024

SOURCE: SARPAS

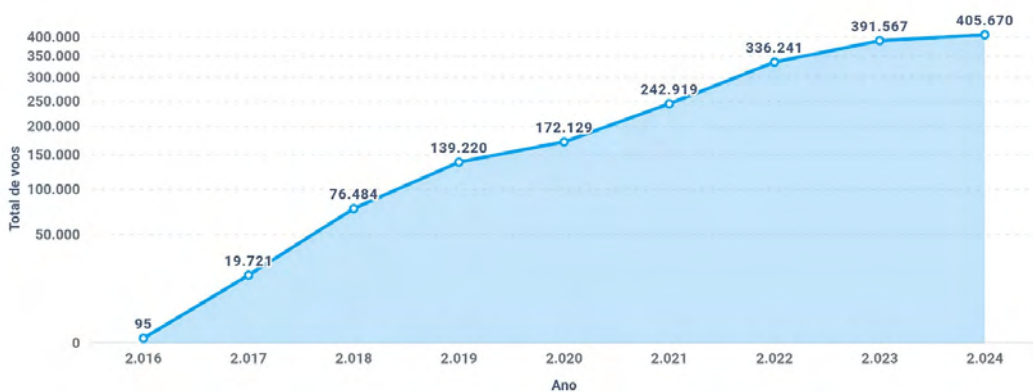


Figure 2 – Evolution of drone flight requests in Brazil between 2020–2024. Source: Adapted from (1).

Taxonomies of UAV missions organize expected deliverables and constraints by output type: orthomosaics, dense 3D models, thermal maps, vegetation indices and by

environment urban, rural, forested or industrial clarifying that different tasks impose distinct requirements on spatial resolution, radiometric fidelity and temporal cadence (8, 21). Early applications in cultural heritage emphasized integrated visible and infrared imaging for high-fidelity reconstruction of façades and architectural elements (18).

Contemporary surveys identify several dominant mission classes: thermal inspection of buildings and photovoltaic arrays, urban cadastral mapping, high-fidelity 3D reconstruction of heritage sites, high-resolution topographic surveys, and multitemporal environmental monitoring. Each demands different balances of geometric accuracy, radiometric calibration, and operational cadence (21, 23, 8, 9, 24). These use cases motivate multimodal sensing designs where RGB, thermal, and LiDAR modalities are planned jointly rather than acquired independently.

Operational rationales for UAV deployment are often summarized by the “three D’s”: dull, dangerous and dirty tasks that are repetitive, hazardous or unsuitable for human presence (25). In such contexts UAVs act as mobile sensing agents capable of repeated, targeted observations and of integrating heterogeneous measurements (RGB, thermal and LiDAR) on a single platform or across coordinated fleets.

Realizing robust multimodal UAV systems requires careful attention to sensor integration, where multimodal refers to the joint use of heterogeneous sensors (RGB cameras, thermal imagers, and LiDAR scanners) mounted on the same aerial platform. This integration demands mechanical mounting strategies, intrinsic and extrinsic calibration, temporal synchronization, and accurate co-registration across modalities. Recent works highlight advances in multimodal alignment between RGB and LiDAR (26), contrastive learning for RGB–thermal fusion (27), and the creation of benchmark datasets combining RGB, thermal, and LiDAR for urban scenarios (28, 29). Systematic reviews of multimodal fusion further emphasize the growing importance of integrated pipelines for UAV-based sensing (30).

Research conducted within the laboratory addresses complementary aspects of these challenges: multi-robot trajectory planning and cooperative coverage strategies inform coordinated multi-platform acquisition (31), architectural proposals for multi-aircraft UAV systems detail hardware–software integration for distributed sensing (32), and IoT-assisted UAV networks exemplify how aerial platforms can augment terrestrial sensor infrastructures for disaster response and resilient monitoring (33).

Motivated by these technological and operational considerations, this thesis proposes a hybrid architecture for three-dimensional reconstruction and feature clustering that integrates RGB, thermal and LiDAR sensing. The central objective is to produce continuous, structured urban models that are robust under realistic mission constraints, that exploit complementary strengths of each modality, and that support practical downstream tasks such as energetic assessment, change detection and asset inspection; the following chapters

detail acquisition strategies, calibration protocols, multimodal data integration techniques and unsupervised feature clustering methods developed and validated in this work.

## 2.2 Three-Dimensional Reconstruction from Images: From Photogrammetry to Computational Geometry

Generating three-dimensional models from two-dimensional images is a well-established practice in photogrammetry and computer vision, repeatedly renewed by advances in sensors, platforms, and algorithms (34, 35).

Historically, photogrammetry relied on calibrated metric cameras and stereoscopic overlaps to derive measurements and maps under controlled conditions. Early approaches emphasized geometric rigor and manual interpretation, producing accurate but labor-intensive reconstructions. The digital transition transformed this discipline into an algorithmic pipeline: collections of overlapping images are processed automatically to estimate camera geometry and scene structure, enabling dense three-dimensional reconstruction in uncontrolled environments (36).

The advent of compact, accurate sensors on UAV platforms intensified this transformation, enabling flexible, high-resolution acquisition for urban mapping, heritage documentation, and infrastructure inspection (8). At the same time, advances in computational geometry and multi-view stereo (MVS) algorithms allowed reconstructions to scale to millions of points and complex scenes, bridging photogrammetry with computer vision.

More recently, deep neural networks (DNNs) have introduced a new paradigm for 3D reconstruction. Learning-based methods employ convolutional and transformer architectures to infer depth, geometry, and semantics directly from image collections, often outperforming classical pipelines in robustness to noise, occlusion, and textureless regions. Techniques such as differentiable rendering, neural radiance fields (NeRF), and Gaussian-based splatting (2, 37, 38) exemplify this shift, enabling continuous, gradient-compatible representations that integrate seamlessly with modern learning frameworks. These approaches extend reconstruction beyond geometry, incorporating semantic understanding and multimodal integration, and thus set the stage for the hybrid architectures explored in this thesis.

This evolution is tightly linked to the broader convergence between image processing, computational geometry and graphical representation. Foundational contributions in image-based modeling and vision-driven reconstruction have expanded the scope of photogrammetric techniques beyond traditional surveying. Works such as *Image Processing for Computer Graphics and Vision* (39), together with seminal IEEE contributions on structure-from-motion (36) and multi-view stereo (40), formalize the mathematical

and algorithmic principles that underpin feature extraction, stereo matching and depth estimation, which remain core components of modern reconstruction pipelines.

More recently, the integration of artificial intelligence with immersive media has opened new directions for spatial representation and semantic modeling. The concept of expanded reality, as explored in *Expanded Reality: New Media and AI* (41), emphasizes the fusion of multimodal data and computational perception to generate enriched, interactive 3D environments. These frameworks extend the utility of photogrammetric reconstruction into domains such as urban simulation, cultural heritage visualization and intelligent infrastructure monitoring.

In this research, these principles are operationalized through a multimodal reconstruction pipeline that combines RGB imagery, thermal data and LiDAR scans. The methodology builds on Structure-from-Motion and Multi-View Stereo techniques to generate dense point clouds, which are then integrated, segmented and rendered using differentiable representations. The resulting models reflect both geometric fidelity and radiometric richness, supporting downstream tasks such as feature clustering and anomaly detection.

### 2.2.1 Digital Photogrammetry and Structure-from-Motion (SfM)

Structure-from-Motion (SfM) constitutes the methodological core of image-based three-dimensional reconstruction for UAV missions (36, 34). SfM jointly estimates camera intrinsics and extrinsics together with a sparse three-dimensional structure from collections of overlapping images. In operational terms, the pipeline begins with image ingestion and preprocessing, proceeds through feature detection and robust matching, performs relative pose estimation and triangulation to obtain an initial sparse point cloud, and finally performs global refinement via bundle adjustment (42, 43). Modern implementations complement this core with Multi-View Stereo (MVS), which densifies the reconstruction by estimating depth across multiple overlapping views, followed by texture mapping, radiometric corrections, and export routines that produce georeferenced point clouds and meshes suitable for subsequent analysis and fusion with other sensors (44). Figure 3 schematizes the conceptual pipeline and shows how theoretical stages map to concrete software modules.

Image undistortion and radiometric pre-processing typically precede feature extraction, where local descriptors such as SIFT (Scale-Invariant Feature Transform) (45), SURF (Speeded-Up Robust Features) (46), and ORB (Oriented FAST and Rotated BRIEF) (47) are commonly used. Matching strategies combine descriptor similarity with geometric verification, including epipolar constraints and RANSAC (Random Sample Consensus)-based outlier rejection (48). The sparse reconstruction obtained by triangulation provides the scaffold for later densification.

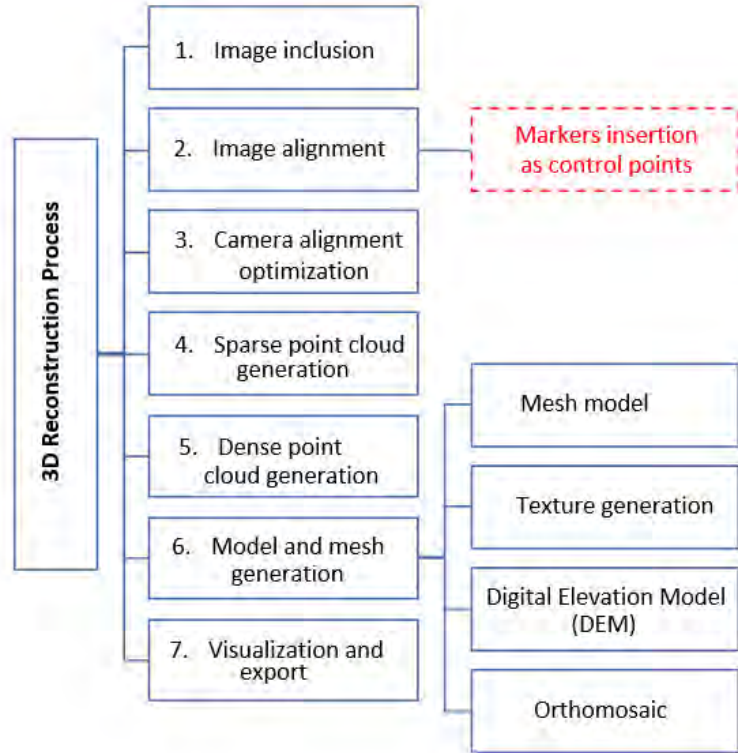


Figure 3 – SfM–MVS reconstruction pipeline with images.

Practical software differs in automation and tuning: commercial packages such as Agisoft Metashape prioritize end-to-end usability and GIS (Geographic Information Systems) integration (44), while research systems like COLMAP expose modular components and parameters that facilitate experimental work (36). Open-source stacks such as OpenSfM and OpenDroneMap increase accessibility for practitioners and researchers, though they often demand careful parameter adjustment in complex urban scenes (49, 50, 51).

### 2.2.1.1 Projection model and notation

The geometric relation between a three-dimensional world point and its image measurement is expressed by the pinhole camera model, as illustrated in Figure 4. In homogeneous coordinates, the projection of point  $\mathbf{X}_i$  onto image  $k$  is written as Equation 2.1:

$$\tilde{\mathbf{u}}_{ik} = \mathbf{K}_k [\mathbf{R}_k \mid \mathbf{t}_k] \mathbf{X}_i \quad (2.1)$$

where  $\mathbf{X}_i = [x_i, y_i, z_i, 1]^\top$  denotes the homogeneous world coordinate,  $\mathbf{K}_k$  is the camera intrinsic matrix and  $[\mathbf{R}_k \mid \mathbf{t}_k]$  is the extrinsic block composed by rotation  $\mathbf{R}_k \in SO(3)$  and translation  $\mathbf{t}_k \in \mathbb{R}^3$  (42). The intrinsic matrix is commonly parameterized as Equation 2.2:

$$\mathbf{K}_k = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (2.2)$$

With  $f_x, f_y$  the focal lengths (in pixels) and  $(c_x, c_y)$  the principal point coordinates. The inhomogeneous image coordinate  $\mathbf{u}_{ik} \in \mathbb{R}^2$  is obtained by normalizing the homogeneous projection  $\tilde{\mathbf{u}}_{ik}$  by its third component.

Conceptually, projection is a two-step mapping (see Figure 4). First, the extrinsic transformation maps the world point into the camera frame:

$$\mathbf{X}_{ik}^{(\text{cam})} = \mathbf{R}_k \mathbf{X}_i + \mathbf{t}_k \quad (2.3)$$

and second, the intrinsics map the camera-frame vector onto the image plane:

$$\tilde{\mathbf{u}}_{ik} = \mathbf{K}_k \mathbf{X}_{ik}^{(\text{cam})} \quad (2.4)$$

Precise modeling of lens distortion and sensor-specific parameters is essential because small calibration errors in  $\mathbf{K}_k$  or misestimations in  $[\mathbf{R}_k \mid \mathbf{t}_k]$  are amplified during triangulation and densification, producing biased geometries (43).

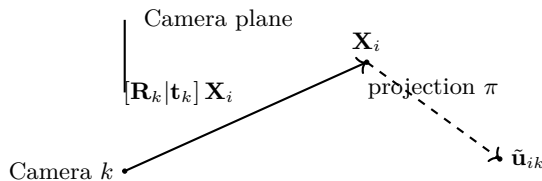


Figure 4 – Pinhole projection model: mapping of a world point  $\mathbf{X}_i$  to its image measurement  $\tilde{\mathbf{u}}_{ik}$ .

### 2.2.1.2 Triangulation and sparse reconstruction

Triangulation recovers a three-dimensional point by intersecting back-projected rays from image observations (42). Given multiple calibrated views of the same feature, each image measurement defines a ray in space; triangulation seeks the point  $\mathbf{X}_i$  that minimizes reprojection residuals across these views. The numerical stability of triangulation depends on baseline geometry, image overlap, the accuracy of feature localization and the angular separation of views: near-collinear camera configurations and small baselines produce ill-conditioned depth estimates and large uncertainty in reconstructed depth (52). This process is illustrated in Figure 5.

### 2.2.1.3 Bundle adjustment and global refinement

Bundle adjustment performs nonlinear optimization to jointly refine camera poses and 3D point positions by minimizing reprojection error across all observations (53). The

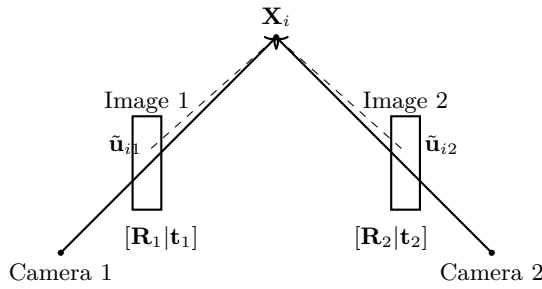


Figure 5 – Triangulation principle: a 3D point  $\mathbf{X}_i$  is recovered by intersecting rays back-projected from multiple calibrated camera views.

canonical objective is:

$$\min_{\{R_i, t_i\}, \{X_j\}} \sum_{i,j} \left\| \mathbf{u}_{ij} - \pi(R_i X_j + t_i) \right\|^2 \quad (2.5)$$

Where  $\mathbf{u}_{ij}$  denotes the observed image coordinate of point  $X_j$  in camera  $i$  and  $\pi(\cdot)$  denotes the projection operator including intrinsics and distortion when modeled. Bundle adjustment implementations exploit visibility sparsity and use block-sparse solvers; for large UAV datasets, hierarchical and constrained variants reduce computational load while maintaining accuracy (54, 55). Robust estimation strategies employ M-estimators or observation weighting to mitigate the impact of outliers from repetitive textures or occlusions.

## 2.2.2 Multi-View Stereo: Densifying structure

While SfM yields a sparse metric scaffold, MVS estimates dense depth and reconstructs detailed surface samples by leveraging photometric consistency across calibrated views (56, 57). MVS formulates per-pixel or per-patch depth estimation as an energy minimization balancing photometric fidelity with spatial regularization, or as probabilistic inference combining multi-view likelihoods and priors. Patch-based, depth-map fusion and volumetric occupancy methods represent different practical paradigms, each with trade-offs between local detail, robustness to noise and computational cost.

Urban scenes present particular challenges for MVS. Specular façades, transparent glazing and reflective materials violate near-Lambertian assumptions, producing spurious matches and erroneous depths. Vegetation and repetitive textures create ambiguous correspondences, while illumination variations and cast shadows complicate photometric comparisons across passes. Mitigation strategies include acquisition planning to maximize uniform illumination and parallax diversity, the use of robust photometric metrics or learned matching functions, and the incorporation of geometric priors or auxiliary LiDAR returns to regularize depth estimation (58).

From an operational perspective, acquisition parameters strongly influence MVS outcomes. Recommended forward overlap values for façade-oriented missions commonly

exceed seventy percent with sidelpap above sixty percent to ensure redundant angular coverage and robust angular diversity. Baseline selection must balance detail and stability: small baselines favor high-frequency detail but reduce depth precision at distance, whereas larger baselines improve depth precision but increase occlusion risk. For building-scale surveys, parallax angles in the range of approximately five to twenty-five degrees often produce a practical compromise between matching stability and depth resolution. Depth sampling resolution in plane-sweep or volumetric methods should reflect camera resolution and scene scale, with finer sampling improving detail at the expense of runtime and noise sensitivity.

Empirical evaluations indicate that well-designed SfM–MVS workflows with precise calibration and adequate overlap can achieve high point densities and sub-decimeter accuracy on vertical façades under favorable conditions, while degraded illumination, low texture or complex materials reduce effective density and accuracy (58). These limitations motivate multimodal data integration with LiDAR and thermal sensing to recover geometry and radiometric attributes in challenging urban contexts.

### Post-processing, meshing and attribute transfer

Post-processing after densification typically includes outlier filtering, normal estimation, mesh reconstruction and attribute mapping. Statistical filtering and normal-consistency checks remove inconsistent samples, while Poisson surface reconstruction methods convert point samples into manifold surfaces suited for visualization and GIS export. In particular, the classical Poisson reconstruction formulates surface extraction as the solution of a spatial Poisson equation, where the gradient of an indicator function is approximated by the input point normals (59). This approach yields watertight surfaces and is robust to noise, making it widely adopted in photogrammetry and computer vision. Screened Poisson reconstruction extends the method by introducing a data-fitting term that balances smoothness with fidelity to the original samples, improving detail preservation in complex geometries (60).

When thermal or LiDAR data are available, radiometric or intensity attributes can be projected onto dense geometry after careful radiometric calibration and accurate geometric co-registration, enabling energetic analyses and material diagnostics on façades and roofs.

### Software pipelines and operational considerations

Software implementations operationalize the SfM–MVS conceptual flow with different trade-offs that directly impact usability, accuracy, and scalability. Commercial solutions such as Agisoft Metashape and Pix4D emphasize automation, end-to-end workflows, and GIS interoperability, offering user-friendly interfaces and integrated georeferencing modules

(44). Their practical limitations include restricted transparency of internal algorithms, licensing costs, and reduced flexibility for experimental configurations. These constraints make them highly effective for production environments but less suited for methodological innovation.

Research-oriented systems such as COLMAP (36) and VisualSfM expose modular components for feature extraction, matching, incremental reconstruction, and bundle adjustment. These modules are highly customizable, allowing researchers to test alternative descriptors, matching strategies, or optimization routines. The trade-off is that such systems demand greater expertise in parameter tuning and often require more computational resources, but they provide reproducibility and extensibility that are essential for scientific work.

Open-source stacks such as OpenSfM, MicMac, and OpenDroneMap (49, 50, 51) target accessibility and community-driven development. They lower entry barriers by enabling UAV practitioners to process data without expensive licenses, but their performance in large-scale or complex urban scenes can be sensitive to parameter choices and hardware constraints. Moreover, support and documentation may vary, requiring users to rely on community forums and shared experience.

Sensor metadata, including RTK GNSS positions, IMU logs and synchronized timestamps, play a critical role in improving initial pose estimates and reducing dependence on purely visual matches in feature-poor regions. Integrating these metadata streams into the pipeline reduces drift, accelerates convergence and enhances robustness in challenging environments such as repetitive façades or vegetation.

Operational workflows commonly combine quick on-site reconstructions for quality assessment with high-fidelity offline processing to produce final metric products. The implications of these trade-offs are significant: commercial packages minimize operator time but may sacrifice algorithmic transparency; research systems maximize methodological control but increase processing time and complexity; open-source stacks democratize access but require careful calibration and parameter tuning to achieve reliable accuracy.

In summary, digital photogrammetry combining SfM and MVS is a flexible and powerful framework for UAV-based 3D reconstruction. Its reliability, however, is conditioned on acquisition design, calibration quality, scene properties, and processing choices. These constraints justify the multimodal approach adopted in this work, which integrates LiDAR and thermal sensing to enhance geometric fidelity and radiometric expressiveness in urban-scale modeling.

## 2.3 Multimodal Integration of Three-Dimensional Data: Uniting Geometry, Color and Temperature

Multisensor data integration for urban three-dimensional mapping pursues a single, coherent representation that is simultaneously metric, photometric and radiometric, enabling robust measurement, diagnostic interpretation and analysis in complex built environments. Cities present a combination of heterogeneous materials, repeating architectural motifs, vegetated occlusions and dynamic elements that expose the limitations of single-modality pipelines: RGB imagery captures high-frequency texture and color needed for discrimination but degrades under poor illumination and on textureless façades; thermal imaging reveals energetic phenomena, heat losses and hotspots invisible in the optical band yet typically exhibits lower spatial resolution and emissivity-dependent radiometry; LiDAR provides metric geometry and surface orientation independent of lighting but usually lacks dense radiometric detail. Deliberate fusion reduces modality-specific uncertainty, closes spatial and radiometric gaps, and produces enriched per-point descriptions that materially improve downstream tasks such as anomaly detection, change monitoring, automated inspection and energy assessment (61, 62, 63, 64).

### 2.3.1 Preprocessing, calibration and shared reference frames

A reliable multimodal product depends on rigorous, modality-specific preprocessing that enforces a common geometric and radiometric frame. LiDAR sensors commonly deliver returns in sensor-centric spherical coordinates  $(r, \theta, \phi)$ ; converting these returns to Cartesian coordinates via Equations (2.6)–(2.8) embeds all geometric observations in a Euclidean scaffold appropriate for neighborhood queries, metric weighting and geometric reasoning:

$$x = r \sin(\phi) \cos(\theta), \quad (2.6)$$

$$y = r \sin(\phi) \sin(\theta), \quad (2.7)$$

$$z = r \cos(\phi). \quad (2.8)$$

These Cartesian points are initially expressed in the local LiDAR reference frame. To integrate them into a mapping coordinate system, a rigid-body transformation is applied, combining rotation and translation derived from sensor calibration and platform navigation data (GNSS/INS). This step ensures that LiDAR-derived geometry is expressed consistently with other modalities and can be directly compared or integrated with photogrammetric reconstructions.

Photogrammetric pipelines then convert RGB image sequences into metric point clouds using Structure from Motion and dense Multi-View Stereo. These pipelines correct intrinsics, model lens distortion and perform bundle adjustment so that forward and back

projection are metrically consistent across views (42, 65, 56). By aligning LiDAR geometry with the mapping frame before introducing RGB-based reconstructions, both modalities share a common spatial reference, which facilitates joint processing and fusion.

Thermal sensors require radiometric calibration that maps raw digital counts into radiance or brightness temperature while compensating for sensor offset, gain, atmospheric path effects and surface emissivity. Without radiometric correction, thermal observations are not interoperable across viewpoints or with in-situ measurements and cannot be trusted for quantitative assessment (63, 64).

Preprocessing therefore includes geometric conversion, photogrammetric correction, radiometric calibration and metadata harmonization (timestamps, camera poses, exposure logs, per-image confidence). High-quality metadata is critical: accurate timestamps, per-image confidence measures, camera temperature logs and exposure parameters directly influence visibility checks, selection of preferred observations and temporal filtering. Where available, absolute georeferencing (GNSS/INS, ground control points) is integrated early to reduce drift and provide pose priors that greatly simplify large-scale co-registration (62, 66).

### 2.3.2 Geometric co-registration: algorithms, consistency and diagnostics

Co-registration of modality-specific point sets is most often modeled as rigid registration since urban scenes are quasi-static at the acquisition time of UAV and mobile surveys. Iterative Closest Point (ICP) and robust variants formulate rigid alignment as the search for a transformation  $T \in SE(3)$  minimizing squared residuals between matched samples:

$$T^* = \arg \min_{T \in SE(3)} \sum_{i=1}^N \|T(p_i) - \ell_{\pi(i)}\|^2, \quad (2.9)$$

where  $p_i$  are source points,  $\ell_{\pi(i)}$  their matched counterparts under correspondence map  $\pi$ , and  $T(p) = Rp + t$  decomposes into rotation  $R \in SO(3)$  and translation  $t \in \mathbb{R}^3$  (67, 68). Practical city-scale implementations augment the canonical ICP loop with hierarchical coarse-to-fine scheduling, voxel/octree decimation, robust correspondence pruning and fast nearest-neighbor indices to ensure tractability.

Urban geometry poses specific challenges: repetitive façades, planar roofs, vegetation and reflective surfaces create ambiguous or spurious correspondences and generate local minima for iterative solvers. Because ICP requires a good initial approximation, our solution was to combine pose priors from GNSS/INS with coarse visual alignment, ensuring that the initial transformation was sufficiently close to the optimum to allow convergence. This initialization step reduced the risk of ICP getting trapped in local minima and improved stability in heterogeneous urban datasets. We further applied per-point weighting schemes that use LiDAR intensity, local sampling density, RGB feature-match confidence or thermal

variance to downweight unreliable matches, and robust loss functions (Huber, Tukey biweight) to limit the influence of gross outliers (69, 70). Attribute-aware registration, where radiometric similarity enters correspondence scoring, often improves alignment stability across heterogeneous sensors and helps preserve architectural detail in the registered scaffold (61).

Diagnostic visualization remains indispensable in practical workflows. Rendered diagnostics that show initial poses, matched correspondences, residual vectors and outlier distributions (Figure 6) complement quantitative residual statistics and guide selective local refinement, reweighting or manual intervention before attribute transfer. Similar strategies have been reported in multimodal registration studies, such as LiDAR–RGB fusion for urban mapping (71), UAV-based thermal–RGB alignment (64), and large-scale city modeling with hybrid ICP pipelines (62, 66). These works reinforce that robust initialization, attribute-aware correspondence and diagnostic inspection are critical to achieving reliable co-registration in complex urban environments.

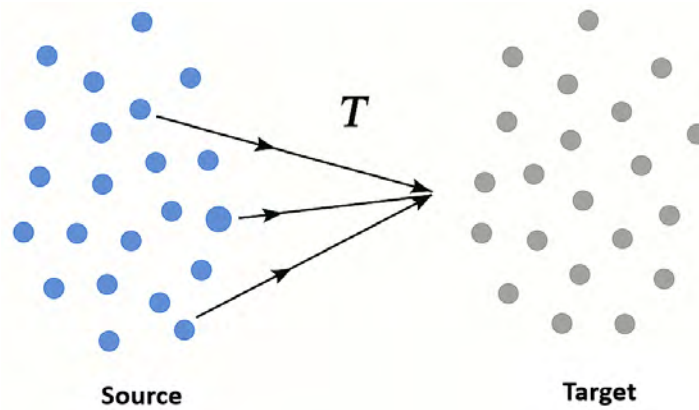


Figure 6 – Geometric alignment example illustrating ICP-based registration between source and target point clouds, showing initial pose, correspondence indicators and final fit for diagnostic inspection.

### 2.3.3 Attribute transfer: projection, Gaussian interpolation and hybrid strategy

Once geometric consistency is reached, attribute transfer populates the unified point set with photometric and radiometric channels (RGB, temperature, intensity). Direct projection of calibrated image pixels to visible 3D points is the preferred method when intrinsics and extrinsics are accurate and visibility tests pass, because it preserves the native sampling, high-frequency radiometric detail and per-image confidence semantics (56, 68). Direct projection, however, depends on three simultaneous conditions: accurate calibration, confirmed visibility (occlusion tests) and temporal coherence. Failing any of these leads to mixed pixels, erroneous assignments or temporal inconsistencies.

Where projection is infeasible, due to occlusions, partial overlap, large incidence angles or sampling density mismatch, a kernel-based interpolation constructs continuous

attribute fields from nearby observations. We employ a Gaussian-weighted estimator that is used both as a fusion primitive and as a principled input to the Gaussian splatting renderer:

$$\hat{a}(x) = \frac{\sum_{i=1}^k a_i \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)}{\sum_{i=1}^k \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)}, \quad (2.10)$$

where  $\{(x_i, a_i)\}_{i=1}^k$  are the  $k$  nearest observations,  $k$  the neighborhood size and  $\sigma$  the kernel bandwidth controlling spatial influence. The numerator aggregates attributes with spatially decaying weights, while the denominator normalizes sums to avoid attenuation. Small  $\sigma$  favors locality and edge fidelity; large  $\sigma$  yields smoother fields but risks blurring edges and erasing small thermal features (72). Weighted interpolation improves consistency compared to simple nearest-neighbor assignment, since points closer to the query contribute more strongly, reducing noise and stabilizing attribute transfer in heterogeneous sampling. In practice, the computational overhead of weighting is modest: once  $k$  nearest neighbors are retrieved, the exponential weights are computed in vectorized form or on GPU, adding less than 10–15% to runtime compared to unweighted assignment in our city-scale tests. The benefit is significant, as weighted interpolation reduces ghosting across façades, preserves thermal contrasts and avoids abrupt attribute jumps that occur when only the closest neighbor is used.

Selecting  $k$  and  $\sigma$  robustly is essential. In our experiments, typical heuristics are  $k = 8\text{--}16$  for dense photogrammetric clouds and  $k = 8\text{--}12$  for sparser LiDAR–thermal combinations. The kernel bandwidth  $\sigma$  is set adaptively as a fraction (0.4–1.0) of the median local  $k$ -nearest neighbor ( $k$ -NN) distance. Contributions beyond  $r_{\max} \approx 3\sigma$  are truncated for efficiency and to avoid numerical underflow. Denominators below a small threshold trigger fallbacks to nearest-neighbor assignment or local  $\sigma$  inflation. Visibility-limited points preferentially use observations with smaller incidence angle or higher camera confidence before interpolation, favoring sharper direct measures when available. The  $k$ -NN search itself identifies the  $k$  closest points to a query in Euclidean space, and efficient retrieval is performed using  $k$ -d trees or octrees, which scale well to millions of points. In anisotropic sampling regions, Mahalanobis distances derived from local covariance are used instead of Euclidean distances to better capture directional support.

Because isotropic Gaussian kernels blur material boundaries, we employ attribute-aware bilateral variants in sensitive regions. In this case, the spatial Gaussian is multiplied by a radiometric similarity term so that neighbors with large color or temperature differences contribute less:

$$w(x, x_i) = \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{\Delta_a(x, x_i)^2}{2\sigma_a^2}\right), \quad (2.11)$$

where  $\Delta_a$  is the radiometric difference and  $\sigma_a$  the radiometric bandwidth. This reduces ghosting across façades and preserves thermal contrasts essential to diagnostics (72, 63). Bilateral filtering has been widely studied in image processing (73) and has proven effective

in point cloud attribute transfer, particularly in multimodal integration tasks where radiometric discontinuities align with geometric boundaries.

The interpolation stage therefore serves two linked functions: filling coverage gaps left by projection and regularizing noisy measurements to deliver stable priors for downstream semantic and energetic analyses. Practical failure modes include overly sparse neighborhoods that force over-smoothing or inventing detail, and strong anisotropic sampling where Euclidean distances misrepresent support. We mitigate these by using adaptive  $\sigma$ , truncation radii, Mahalanobis distances derived from local covariance when sampling is anisotropic, and by falling back to nearest neighbors or marking points as unassigned when data are insufficient. Computationally, efficient k-NN retrieval (tiled k-d trees, octrees) and vectorized or GPU weight computation are used to scale interpolation to city-scale clouds. Pre-filtering by timestamp and per-image confidence excludes temporally inconsistent or low-quality observations prior to interpolation. When multiple observations project to the same pixel, selection rules prefer lower incidence angles or higher confidence measurements to reduce projection noise. Similar strategies have been reported in multimodal integration studies, including thermal–RGB UAV mapping (64), LiDAR–RGB fusion for urban modeling (71), and bilateral filtering in point cloud denoising (74), reinforcing the benefits of weighted and bilateral interpolation in complex urban environments.

## Connection to Gaussian splatting and rendering

Gaussian splatting is a rendering technique where each 3D sample is represented not as a discrete point but as a small Gaussian “splat” projected into the image plane. Each splat has a footprint defined by its radius and covariance, which control how much influence it has in space and how anisotropic its support can be. Radiometric channels such as RGB color or thermal intensity are attached to each splat, so that rendering produces smooth images that reflect both geometry and attributes.

The Gaussian interpolation formalism described earlier naturally bridges fusion and rendering. Local bandwidth  $\sigma$  and neighborhood covariance provide a principled mapping to per-splat radii and anisotropic covariance matrices, while interpolated radiometric channels (RGB, temperature) supply splat color and thermal attributes. Maintaining consistency between the interpolation kernel support and the splat footprint ensures that the rendered appearance reflects the same statistical support used during fusion. This prevents visual–analytic mismatches and produces smoother, more reliable renderings for inspection and validation. Moreover, attribute-aware interpolated values reduce high-frequency radiometric noise that would otherwise cause flicker or aliasing in splatted views.

In the context of this thesis, Gaussian-weighted interpolation was applied selectively to fill gaps left by direct projection, particularly in regions of partial occlusion or sparse

LiDAR–thermal sampling. Bilateral weighting, which combines spatial proximity with radiometric similarity, was tested in façade regions to preserve thermal contrasts. The fusion pipeline was not fully automated; interpolation was used as a controlled post-processing step to complement direct projection in the multimodal workflow, ensuring that the rendered products remained consistent with the theoretical framework of Gaussian splatting.

### 2.3.4 Serialization, engineering and validation

The integrated, attributed point cloud is serialized in a format that preserves per-point attributes and metadata; we use the Polygon File Format (`.ply`) for its flexibility in storing arbitrary vertex properties (coordinates, RGB, temperature, intensity, labels) and for broad compatibility with visualization (Potree), analysis (CloudCompare) and learning frameworks (PointNet, DGCNN, Point-MAE). For large urban datasets, engineering steps such as stratified sampling, octree compression, attribute quantization and metadata packaging are necessary to keep I/O practical while preserving analytic fidelity.

Operational constraints inform acquisition and processing. Thermal images usually have lower native resolution and narrower fields of view than RGB cameras; harmonization strategies include multi-altitude passes, super-resolution approaches and multi-camera arrays to reduce the mismatch in scale. Differing viewing geometries cause parallax and occlusion; occlusion-aware projection, mesh-based visibility testing and multi-view fusion heuristics are applied to avoid incorrect assignments. Temporal misalignment requires synchronization or timestamp-based filtering before interpolation (63, 64). To extend these registration and fusion strategies beyond small test cases and into practical urban deployments, scalability becomes a central requirement. Algorithmic scalability relies on hierarchical ICP, efficient nearest-neighbor structures, GPU-accelerated fusion kernels and compressed storage to make the pipeline applicable to city-scale problems (56, 68).

Empirical evaluations support the utility of multimodal fusion: integrated UAV RGB, thermal and LiDAR pipelines improve detection of thermal bridges and rooftop defects, produce more reliable building envelope energy assessments and yield richer façade condition maps than single modality products, with measurable gains in detection sensitivity and spatial localization (64, 75, 58). These findings underpin the methodological choices described here, including rigorous calibration, attribute-aware weighted registration (Equation 2.9), adaptive Gaussian interpolation (Equation 2.10) and conservative post-filtering, ensuring that the integrated models are robust inputs to segmentation, energy assessment and operational monitoring.

In summary, multimodal integration for urban 3D modeling is a multi-stage, tightly integrated process that interleaves precise sensor preprocessing, robust rigid registration and adaptive attribute transfer. When combined with considered acquisition design and

scalable computation, the integrated models merge LiDAR’s metric fidelity, RGB’s semantic richness and thermal sensing’s energetic insight, producing actionable urban representations for monitoring, diagnostics and asset management (61, 64, 63, 66).

## 2.4 Continuous Rendering with Gaussian Splatting

Gaussian Splatting introduces a continuous and differentiable rendering paradigm that replaces discrete point primitives with spatially extended, anisotropic Gaussian functions. This approach redefines the representation of 3D scenes by modeling each point not as a singular location but as a probabilistic density with spatial support, orientation and opacity. The result is a smooth, analytically tractable formulation that enables high-fidelity visualization and gradient-based optimization, bridging the gap between traditional geometric modeling and modern neural rendering techniques.

Figure 7 illustrates the Gaussian Splatting pipeline proposed by (2), which converts an initial SfM reconstruction into a continuous differentiable representation that can be used by deep learning methods.

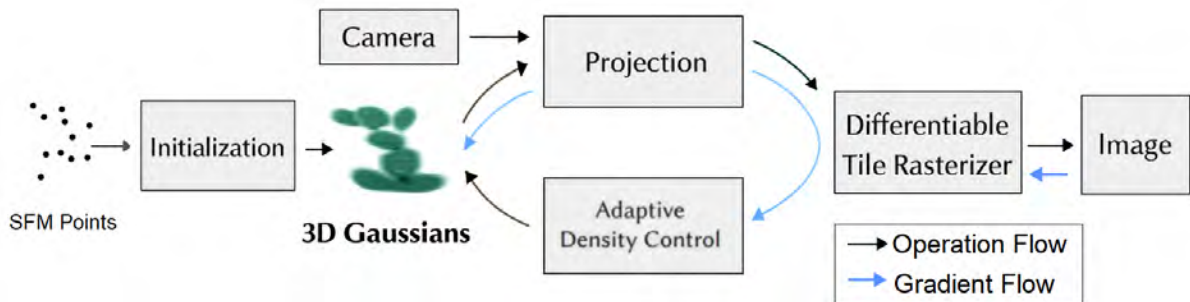


Figure 7 – Operational flow of the continuous representation module using Gaussian Splatting from SfM reconstructions. Source: (2).

Pipeline of Gaussian Splatting’s stages:

1. **SfM points:** reconstruct initial sparse geometry  $\mathcal{P} = \{\mathbf{p}_i\}$ .
2. **Initialization:** convert each  $\mathbf{p}_i$  into an initial Gaussian with  $\mu_i = \mathbf{p}_i$  and  $\Sigma_i = \sigma^2 I$ , and assign attributes.
3. **Camera projection:** project Gaussians into the image plane via

$$\mathbf{u}_i = \mathbf{K} \cdot (\mathbf{R} \cdot \mu_i + \mathbf{t}), \quad (2.12)$$

linking continuous scene representation to 2D visualization.

4. **Adaptive density control:** during training, adjust covariance  $\Sigma_i$ , opacity  $\alpha_i$  and orientation to ensure uniform coverage and avoid redundancy or holes.

5. **Differentiable Tile rasterizer:** integrate Gaussian densities over image space into pixels using soft z-buffering to handle occlusion smoothly; gradients propagate for end-to-end training.
6. **Final image:** produce a continuous image combining synthetic color and depth coherent with scene geometry, usable for downstream feature clustering and structural analysis.

This pipeline transforms sparse discrete reconstructions into continuous, differentiable representations that capture urban environmental detail while remaining compatible with modern learning techniques (2, 72).

At the core of the method lies the representation of each splat  $G_i$  as a Gaussian function defined by a center  $\mu_i \in \mathbb{R}^3$ , a covariance matrix  $\Sigma_i \in \mathbb{R}^{3 \times 3}$ , and an opacity coefficient  $\alpha_i \in [0, 1]$ . The density function is given by:

$$G_i(x) = \alpha_i \cdot \exp\left(- (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right), \quad (2.13)$$

This formulation captures the anisotropic spread of each splat through the Mahalanobis distance, allowing the shape and orientation of the splat to reflect local geometric structure. The opacity  $\alpha_i$  modulates the visual contribution of each splat, enabling soft blending and occlusion-aware compositing. When  $\Sigma_i = \sigma^2 I$ , the splat is isotropic, and its influence decays uniformly in all directions. In more general cases,  $\Sigma_i$  encodes directional uncertainty, elongation along surface tangents, and compression along normals, making the representation sensitive to local surface geometry.

The projection of splats into the image plane is performed using standard camera models. Given a camera with intrinsic matrix  $K$  and extrinsic parameters  $(R, t)$ , the center of the splat is projected as:

$$\mathbf{u}_i = K \cdot (R \cdot \mu_i + t), \quad (2.14)$$

The covariance  $\Sigma_i$  is propagated through the Jacobian of the projection function to yield a 2D elliptical footprint in image space. This footprint defines the region over which the splat contributes to pixel intensities. Crucially, both the projected center and the footprint are differentiable with respect to the splat parameters, enabling gradient flow from image-space losses back to the 3D representation.

Figure 8 illustrates the geometric intuition behind this formulation. The splat is shown as an ellipsoidal volume in 3D space, with its center  $\mu_i$ , orientation and extent defined by  $\Sigma_i$ . The projection yields a 2D ellipse in the image plane, which determines the splat's contribution to the rendered image.

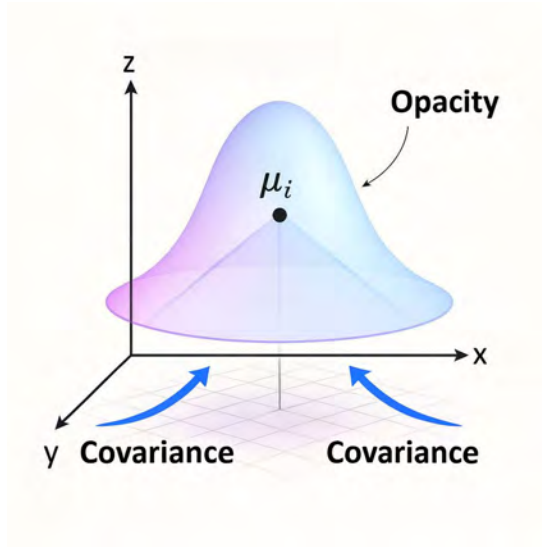


Figure 8 – Geometric interpretation of an anisotropic Gaussian splat in 3D space. The mean  $\mu_i$  defines the center of the distribution, while the covariance governs its spread and orientation along the axes. Opacity modulates the visual contribution of the splat, illustrated here by the semi-transparent surface.

To ensure geometric completeness and radiometric richness from the outset, the operational pipeline begins with an integrated point cloud obtained from Structure-from-Motion (SfM) and LiDAR data. SfM reconstructions offer dense coverage and high-resolution RGB attributes, while LiDAR contributes metrically accurate geometry, especially in textureless or shadowed regions where photogrammetry fails. This integration combines the strengths of both modalities, producing a more reliable and attribute-rich scaffold for downstream rendering.

Each point in the integrated cloud is converted into a Gaussian splat by assigning its position to  $\mu_i$ , estimating  $\Sigma_i$  from local sampling statistics, and initializing  $\alpha_i$  based on observation confidence. The covariance is typically initialized as isotropic, with  $\sigma$  proportional to the median distance to the  $k$ -nearest neighbors. This ensures that the initial splat support reflects the local density of the point cloud and aligns with the bandwidth used in Gaussian interpolation during integration.

During optimization, the splat parameters are refined to improve coverage, reduce redundancy and enhance visual fidelity. Covariances are adapted to fill gaps in sparse regions and to shrink in well-sampled areas. Opacities are adjusted to suppress splats that contribute little to the rendered image. The differentiable rasterizer integrates the contributions of all splats using soft compositing and occlusion-aware blending. This process accumulates radiometric attributes weighted by the projected Gaussian densities and modulated by depth-based visibility functions. The resulting image is a smooth, continuous rendering that preserves geometric detail and radiometric consistency.

Loss functions defined in image space (photometric reprojection error, thermal consistency, and task-specific reconstruction terms) supply gradients that update the splat parameters. Because the rendering pipeline is differentiable, these gradients propagate through projection and compositing, enabling end-to-end optimization of the continuous scene representation and joint modeling of RGB and thermal channels.

Gaussian Splatting supports multimodal attributes by associating each splat with a vector of radiometric properties. These attributes are blended during rendering using the same spatial weights as the geometric contributions. For thermal data, which often has lower resolution and higher noise, attribute-aware compositing and bilateral filtering are employed to preserve meaningful structure and suppress artifacts. The Gaussian interpolation used during integration provides stable priors for these attributes, ensuring consistency between the integrated point cloud and the rendered image.

Numerical stability is maintained by truncating splat footprints at a fixed multiple of the standard deviation, typically  $3\sigma$ , and by regularizing the covariance matrices to prevent degeneracy. Splats with negligible opacity or redundant support are pruned to reduce computational load. In regions with high residual error, new splats may be introduced or existing ones refined to capture fine detail. In summary, Gaussian Splatting offers a mathematically rigorous and computationally efficient framework for continuous rendering of multimodal 3D scenes. By modeling each point as a spatially extended, differentiable density, it enables smooth visualization, joint optimization and integration with deep learning pipelines. The method complements the integration process by extending the Gaussian formalism from attribute interpolation to rendering, creating a unified analytical structure that supports both reconstruction and analysis.

To prevent saturation in dense regions of the point cloud, opacity contributions are normalized across overlapping splats, ensuring that the cumulative transparency remains bounded. This normalization avoids visual artifacts and preserves contrast in high-density areas. Additionally, splat opacity is modulated by visibility estimates derived from soft z-buffering, which accounts for partial occlusions and depth uncertainty during projection. This mechanism improves rendering realism by attenuating contributions from occluded or low-confidence points, especially in urban scenes with complex geometry.

When applied to urban environments, Gaussian Splatting produces realistic, analyzable models that combine geometric fidelity with radiometric richness, advancing the state of the art in 3D scene understanding and representation.

## 2.5 Unsupervised Feature Clustering in 3D Clouds: Latent Embeddings and Structural Coherence

Feature clustering in three-dimensional point clouds aims to assign consistent labels to spatially coherent regions, enabling structural interpretation and material classification. Unlike raster images, which benefit from regular grid connectivity and convolutional priors, point clouds are unordered sets in  $\mathbb{R}^3$ , often sparse, noisy and lacking explicit topology. These characteristics demand specialized architectures and unsupervised strategies capable of extracting latent representations and discovering groupings without manual annotation.

Let  $\mathcal{P} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^3$  denote a point cloud, where each element  $\mathbf{x}_i \in \mathbb{R}^3$  is a 3D vector representing a point in space. Any function  $f$  operating on  $\mathcal{P}$  must satisfy permutation invariance:

$$f(\{\mathbf{x}_1, \dots, \mathbf{x}_n\}) = f(\{\mathbf{x}_{\sigma(1)}, \dots, \mathbf{x}_{\sigma(n)}\}), \quad (2.15)$$

for any permutation  $\sigma$ . This constraint motivates the use of shared pointwise encoders and symmetric aggregation functions, as pioneered by PointNet (11). In PointNet, each vector  $\mathbf{x}_i$  is passed through a shared multilayer perceptron  $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}^d$ , producing a feature vector in  $\mathbb{R}^d$ . A global feature is then computed via max-pooling:

$$f(\mathcal{P}) = \gamma \left( \text{MAX}_{\mathbf{x}_i \in \mathcal{P}} \phi(\mathbf{x}_i) \right), \quad (2.16)$$

where  $\gamma$  maps the aggregated feature vector to task-specific outputs. While PointNet ensures permutation invariance, it lacks explicit modeling of local geometric relations.

To address this limitation, PointNet++ (76) introduces hierarchical feature extraction. It performs Farthest Point Sampling to select representative centers, defines local neighborhoods via Ball Query:

$$\mathcal{N}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{P} : \|\mathbf{y} - \mathbf{x}\| \leq r\}, \quad (2.17)$$

where  $\mathbf{y}$  and  $\mathbf{x}$  are 3D vectors, and  $r$  is a scalar radius. Localized PointNet modules are then applied to compute features:

$$f_{\text{local}}(\mathbf{x}) = \gamma \left( \text{MAX}_{\mathbf{y} \in \mathcal{N}(\mathbf{x})} \phi(\mathbf{y} - \mathbf{x}) \right), \quad (2.18)$$

where  $\mathbf{y} - \mathbf{x}$  denotes the relative vector between neighbors and the center point, ensuring that local geometric relations are explicitly encoded.

Dynamic Graph Convolutional Neural Network (DGCNN) (13) further enhances local modeling by constructing dynamic graphs in feature space. Each point is connected to its  $k$ -nearest neighbors, and the EdgeConv operator computes:

$$\mathbf{x}'_i = \text{MAX}_{j \in \mathcal{N}(i)} \{ \phi(\mathbf{x}_i, \mathbf{x}_j - \mathbf{x}_i) \}, \quad (2.19)$$

where  $\phi$  is a learnable function (typically a multilayer perceptron). The graph is updated at each layer, allowing adaptive modeling of topological variations such as edges, folds and discontinuities.

Point-MAE (Masked Autoencoder for Point Clouds) (14) introduces transformer-based masked autoencoding. It partitions the cloud into tokens via Farthest Point Sampling, encodes visible tokens with a transformer encoder, and reconstructs masked tokens with a decoder by minimizing the Euclidean reconstruction loss:

$$\mathcal{L} = \|\mathbf{X} - \hat{\mathbf{X}}\|_2^2, \quad (2.20)$$

where  $\mathbf{X}$  denotes the original point cloud (as a matrix of 3D coordinates and attributes) and  $\hat{\mathbf{X}}$  its reconstruction. This self-supervised objective promotes contextual representations that are robust to occlusion and sampling irregularities.

To accelerate neighborhood queries and clustering, spatial indexing structures such as KD-tree (67) are employed. The KD-tree recursively partitions space along the dimension of highest variance using the median:

$$\mu = \text{median}\{x_{d^*} \mid \mathbf{x} \in \mathcal{P}\}, \quad (2.21)$$

reducing query complexity from  $O(n)$  to  $O(\log n)$  in ideal cases. This indexing supports efficient  $k$ -NN retrieval for PointNet++ (an extension of PointNet with hierarchical feature extraction), DGCNN and clustering algorithms.

After extracting latent vectors  $f(p_i) \in \mathbb{R}^{128}$ , unsupervised clustering identifies coherent regions. KMeans partitions the embedding space into  $K$  clusters by minimizing intra-cluster variance:

$$\min_{C_1, \dots, C_K} \sum_{i=1}^K \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|_2^2, \quad (2.22)$$

where  $\mu_i$  is the centroid of cluster  $C_i$ . Density-Based Spatial Clustering of Applications with Noise (DBSCAN) defines clusters based on local density:

$$|\{\mathbf{y} \in \mathcal{P} : \|\mathbf{x} - \mathbf{y}\|_2 \leq \epsilon\}| \geq \text{MinPts}, \quad (2.23)$$

allowing arbitrary-shaped clusters and outlier detection. Hierarchical DBSCAN (HDBSCAN) extends DBSCAN hierarchically, automatically determining the number of clusters and extracting dense regions at multiple scales.

To visualize the latent space and assess clustering quality, dimensionality reduction techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) (77) and Uniform Manifold Approximation and Projection (UMAP) (78) are applied. t-SNE minimizes the Kullback–Leibler divergence between pairwise similarity distributions, preserving local structure. UMAP constructs a fuzzy topological graph and minimizes:

$$L_{\text{UMAP}} = \sum_{i \neq j} w_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|_2^2, \quad (2.24)$$

where  $w_{ij}$  encodes neighborhood strength and  $\|\cdot\|_2$  denotes the Euclidean norm. These projections reveal latent clusters and separability.

Figure 9 shows t-SNE and UMAP projections of the latent space, illustrating how the model organizes points into semantically meaningful regions.

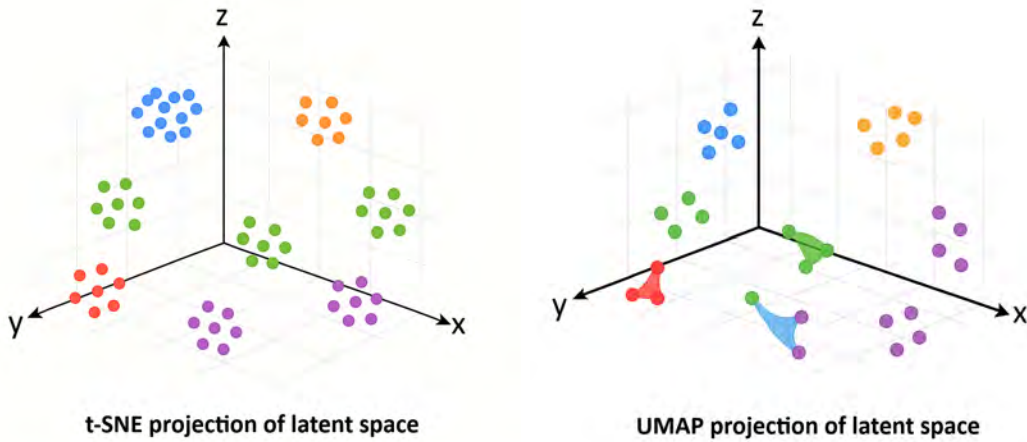


Figure 9 – Projection of latent embeddings using t-SNE (left) and UMAP (right). Each point represents a region in the 3D scene with similar geometric and radiometric properties.

Cluster quality is quantitatively assessed using internal metrics. The silhouette index measures cohesion and separation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (2.25)$$

Where  $a(i)$  is the average intra-cluster distance and  $b(i)$  the nearest-cluster distance. The Davies–Bouldin index evaluates cluster compactness and separation:

$$DB = \frac{1}{K} \sum_{i=1}^K \max_{j \neq i} \left( \frac{\sigma_i + \sigma_j}{\|c_i - c_j\|} \right), \quad (2.26)$$

Where  $\sigma_i$  is the dispersion of cluster  $i$  and  $c_i$  its centroid. Lower DB values indicate better clustering.

Figure 10 illustrates the conceptual foundations of three internal metrics commonly used to evaluate the quality of unsupervised clustering in latent embedding spaces. These visualizations support the interpretation of segmentation results by highlighting how each metric captures distinct aspects of cluster structure—cohesion, separation and geometric fidelity across different architectures and clustering algorithms.

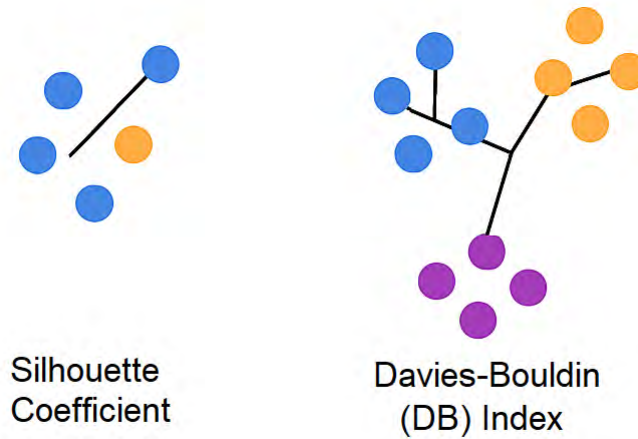


Figure 10 – Conceptual diagrams of clustering evaluation metrics. Left: Silhouette Coefficient, measuring the relative proximity of a point to its own cluster versus neighboring clusters. Right: Davies–Bouldin Index, quantifying intra-cluster dispersion and inter-cluster separation.

To further validate the coherence of the latent clusters, the embeddings were visualized using graph-based connectivity structures. Spatial Neighborhood Trees (SNTs) were constructed over the latent space, revealing how points with similar embeddings form topologically consistent regions. These graphs highlight the ability of the hybrid architecture, combining PointNet++ (76), DGCNN (13) and Point-MAE (14), to capture both local geometry and global context. The resulting clusters align with architectural elements such as façades, rooftops, vegetation and thermal anomalies, even in the absence of explicit supervision.

To capture complementary aspects of 3D structure, three algorithms for feature extraction are employed before projection into a compact embedding. PointNet++ extracts hierarchical features that capture multiscale local geometry by aggregating neighborhoods around representative centers, ensuring sensitivity to both fine and coarse structural patterns. DGCNN, through its EdgeConv operator, extracts topological features that

model point–neighbor relations via dynamic graphs in feature space, learning edge-aware descriptors that respond to boundaries, folds and discontinuities. Point-MAE extracts contextual features by encoding global semantic information using transformer-based masked autoencoding, leveraging reconstruction of masked tokens to learn robust representations under occlusion and sampling irregularities. Together, these algorithms produce feature vectors that represent local detail, relational topology and global context in a complementary manner.

The hybrid embedding function used in this thesis is defined as:

$$f(p_i) = \text{MLP} \left( \text{Concat} \left[ \underbrace{\text{PointNet++}(v_i)}_{\text{hierarchical}}, \underbrace{\text{EdgeConv}(v_i)}_{\text{topological}}, \underbrace{\text{MAE}(v_i)}_{\text{contextual}} \right] \right) \in \mathbb{R}^{128}, \quad (2.27)$$

where  $\mathbf{v}_i$  denotes the input feature vector of point  $p_i$  (including coordinates and multi-modal attributes such as color, temperature and intensity). The concatenated outputs are reduced via a multilayer perceptron to produce a compact embedding. Equation (2.27) integrates multiscale geometry, adaptive connectivity and global context, yielding robust representations for clustering and segmentation.

The clustering algorithms KMeans, DBSCAN and HDBSCAN were applied to these embeddings, and their outputs were evaluated using the metrics described above. HDBSCAN, in particular, demonstrated superior performance in scenes with irregular density and complex topology, automatically determining the number of clusters and identifying outliers. The resulting segmentations were visualized both in latent space and projected back onto the 3D geometry, confirming their spatial coherence and plausibility.

Figure 11 presents a visual summary of the segmentation process, showing the embedding space projections, clustering outputs and the final segmented cloud.

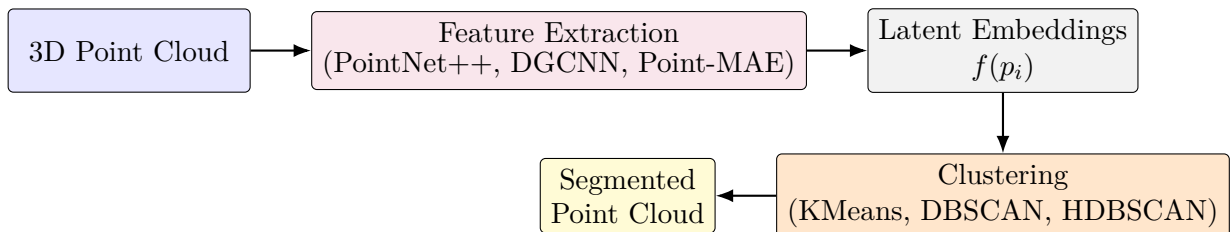


Figure 11 – Simplified conceptual flow of the unsupervised segmentation pipeline. Starting from a 3D point cloud, features are extracted using neural architectures, embeddings are computed, clustering is performed in latent space, and labels are projected back onto the segmented cloud.

The segmentation results demonstrate that unsupervised learning in 3D clouds can achieve meaningful partitioning without manual labels, provided that the embedding architecture captures sufficient geometric and contextual information. The use of multimodal attributes—color, temperature, intensity—further enhances the discriminative power of the embeddings, allowing the clustering algorithms to separate regions not only by shape but also by radiometric behavior.

These findings support the broader thesis that multimodal data integration and continuous representation (via Gaussian-based Splatting (2)) provide a fertile foundation for unsupervised feature clustering segmentation. The latent space learned by the hybrid architecture reflects both structural and energetic properties of the urban environment, and the clustering algorithms reveal coherent groupings that correspond to functional and material distinctions.

## Conclusion

This chapter established the theoretical framework underpinning the methodology proposed in this research. It began by contextualizing the evolution of RPAS platforms and the integration of optical, thermal and LiDAR sensors, highlighting the diagnostic potential of multimodal data acquisition for urban monitoring and structural analysis.

The principles of 3D reconstruction were presented, including photogrammetric techniques such as Structure-from-Motion and Multi-View Stereo, and the mathematical foundations of projection, pose estimation and geometric alignment. Sensor data integration was introduced as a solution to the limitations of individual modalities, with detailed discussion of rigid registration via ICP, attribute transfer through direct projection and Gaussian interpolation, and the use of hybrid strategies to ensure radiometric completeness and edge preservation.

Gaussian Splatting was investigated as a continuous rendering technique that converts discrete point clouds into differentiable scene representations; its mathematical formulation, operational pipeline, and integration with integrated radiometric and geometric attributes were detailed, demonstrating its suitability for visualization, optimization, and downstream analysis.

Finally, the chapter addressed unsupervised segmentation in 3D clouds, detailing the architectures used for latent representation extraction—PointNet, PointNet++, DGCNN and Point-MAE—and the clustering algorithms applied to the resulting embeddings. Spatial indexing via KD-tree, dimensionality reduction via UMAP and t-SNE, and evaluation metrics such as silhouette and Davies–Bouldin were incorporated to assess segmentation quality and coherence.

Together, these components form a comprehensive theoretical foundation for the experimental methodology presented in the next chapter. The concepts and techniques described here support the design and implementation of the `GaussianFusion_AI` pipeline and will be referenced throughout the evaluation and discussion of results.

### 3 INTEGRATED REVIEW OF SCIENTIFIC LITERATURE

Three-dimensional reconstruction of urban environments using data acquired by Remotely Piloted Aerial Systems (RPAS) integrates advances from computer vision, remote sensing, geometric modeling and machine learning. Over the last decade, improvements in sensor miniaturization, platform autonomy, onboard geo-referencing and photogrammetric workflows have enabled systematic acquisition of high-density multimodal datasets, shrinking the gap between experimental demonstrations and operational monitoring. Several contributions by the author have explored these themes in applied contexts, including drone-based 3D modeling of historical buildings (79, 80), reinforcement learning for photogrammetric optimization (79), multi-drone reconstruction of architectural structures (81), and multimodal challenges in urban mapping (82), which provide practical grounding and motivate the architectural choices presented in this thesis. Concurrent developments in representation learning, differentiable rendering and self-supervised pretraining have created new pathways to reduce dependency on manual annotation and to exploit multimodal consistency as a supervisory signal for joint calibration, integration and inference (83, 2, 68). This chapter synthesizes the literature across three interrelated axes — multimodal sensor data integration, continuous neural representations, and unsupervised 3D representation learning — and positions the proposed *GaussianFusion\_AI* architecture within the state of the art (2).

The review process followed a systematic approach to ensure breadth and relevance. Searches were conducted in scientific databases such as IEEE Xplore, SpringerLink, ScienceDirect and platforms made available by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brazil’s Federal Agency for Support and Evaluation of Graduate Education), using combinations of keywords including “*multimodal 3D reconstruction*”, “*Gaussian splatting*”, “*unsupervised point cloud learning*”, and “*photogrammetry LiDAR integration*”. Boolean operators were applied to refine queries, and filters prioritized recent publications, highly cited works and articles in journals of recognized impact. Cross-referencing citations from seminal papers and survey articles helped reduce the risk of omitting relevant contributions. The inclusion criteria focused on works addressing multimodal data integration, continuous neural representations or unsupervised learning in 3D environments, while papers restricted to single modalities or without experimental validation were excluded. This process ensured that the state of the art was represented comprehensively and that the architectural choices of this thesis were grounded in the most relevant advances (83, 2).

A recurring challenge identified in the literature is the scarcity of large-scale multimodal datasets for urban 3D reconstruction. While benchmarks exist for individual

modalities, integrated datasets that jointly provide geometry and radiometric attributes (e.g., RGB and thermal) remain limited in coverage, diversity and annotation depth for complex urban scenarios. Initiatives such as Toronto-3D and SensatUrban broaden the landscape of multimodal resources, yet studies note constraints related to acquisition logistics, sensor synchronization and labeling costs that hinder scalability and generalization across diverse environments (84, 85). This insufficiency motivates the methodological choice of this thesis: to employ Gaussian-based continuous representations that can interpolate across modalities and reduce dependency on exhaustive multimodal corpora, aligning with recent advances in differentiable rendering and neural scene representations (2, 83, 68).

### 3.1 Multimodal Data Integration

Multimodal data integration for urban reconstruction uses complementary strengths of heterogeneous sensors. RGB cameras provide dense high-frequency photometric information useful for texture and fine-detail recovery, LiDAR devices supply direct metric geometry and the ability to penetrate canopy, and thermal sensors encode radiometric measurements associated with material properties and energetic state. The literature documents systematic improvements in completeness and consistency when these modalities are combined, particularly under adverse illumination, occlusion and vegetated scenarios where single-modality pipelines degrade (9, 62). Early surveys and methodological reviews emphasize the central roles of cross-modal calibration, consistent metadata, and principled uncertainty treatment for accurate attribute transfer and semantic enrichment of geometric primitives (86, 87).

Practical integration strategies vary from deterministic geometric projection and attribute rasterization to probabilistic formulations that explicitly model sensor noise, temporal drift and sampling density heterogeneity. Robust geometric co-registration methods combine feature-based alignment, Iterative Closest Point (ICP) variants augmented with radiometric cues, and direct photometric-LiDAR alignment techniques to reduce projection error when mapping thermal or RGB channels onto point clouds (88, 89, 90). Although ICP remains a widely adopted baseline for point cloud registration, its limitations are well documented: sensitivity to initialization, convergence to local minima, and reduced consistency in multimodal contexts where geometric overlap is sparse or radiometric attributes dominate. These constraints motivate the exploration of hybrid approaches that integrate geometric and radiometric information to improve alignment fidelity.

Time synchronization and motion-aware interpolation are particularly important for UAV platforms where thermal cameras can stream at higher frame rates than LiDAR or where sensor sampling epochs differ across passes. Before selecting the adopted strategy, different synchronization methods were evaluated, including hardware-based triggering

between sensors, GPS-disciplined clock alignment, and software-level timestamp interpolation. Hardware triggering ensures deterministic synchronization but is limited by platform integration constraints; GPS-based alignment provides global consistency but suffers from latency and jitter in urban canyons; software interpolation offers flexibility but introduces uncertainty when frame rates diverge significantly. Motion compensation strategies and uncertainty-weighted rasterization, as reported in recent studies (91, 92, 93), were incorporated to mitigate temporal misalignment and preserve per-point attribute fidelity in the integrated datasets.

Beyond low-level registration, higher-level integration architectures explore learned cross-modal attention and transformer based integration blocks that adaptively weight modalities according to context, occlusion patterns and modality reliability. These learned integration layers have shown improved grouping discrimination and consistency to sensor degradation in dense urban scenes (94, 95, 96). Engineering considerations in operational contexts include standardized attribute serialization, provenance-preserving metadata, modular ingestion layers and scalable storage formats that retain per-sample multispectral attributes; toolkits such as Open3D and community specifications for multimodal point clouds are essential for reproducibility but the community still lacks sufficiently large public datasets that jointly provide calibrated RGB, thermal and LiDAR streams with comprehensive ground truth (68, 97, 98, 99). These dataset limitations constrain comparative evaluation and generalization studies, motivating proposals for public benchmarks and comprehensive dataset curation (28, 29).

Operational use cases further expose practical trade-offs. High-fidelity per-point radiometry supports energy mapping, thermal anomaly detection and material classification, but radiometric calibration across different thermal sensors and emissivity conditions is nontrivial and frequently requires scene- or material-specific modeling. Sensor data integration workflows must therefore balance computational cost, calibration complexity and end-user analytic needs, favoring modular pipelines that permit staged processing and incremental refinement as new passes or calibration information become available (100, 101). The literature underscores the value of hybrid operational architectures that mix offline batch reconstruction with streaming-friendly modules to enable near-real-time diagnostics while preserving the ability to perform more expensive global optimizations when computational budget allows (102, 103).

## 3.2 Continuous Representations and Neural Scene Modeling

Continuous neural representations have reshaped the landscape of 3D modeling by offering differentiable, dense functions that compactly encode appearance and geometry. Neural Radiance Fields (NeRF) established a paradigm in which scene radiance and

volumetric density are modeled as continuous functions parameterized by neural networks, enabling high-quality view synthesis and implicit geometry recovery from multi-view images (83). Subsequent engineering efforts addressed NeRF’s scalability limitations, introducing multiresolution hash encodings and other techniques that dramatically reduce memory footprint and accelerate optimization on larger scenes (104). Parallel work introduced Gaussian Splatting, which represents scenes as collections of anisotropic Gaussians or splats that project efficiently and allow gradient-based fitting; Gaussian Splatting offers a compelling trade-off between compactness, real-time rendering capability and direct compatibility with point-cloud-centric processing (2, 105).

Continuous methods have been extended to carry per-sample attributes beyond RGB color, enabling multispectral, thermal and channels to be encoded within the same continuous framework. Multispectral Gaussian encodings and surfel-like Gaussian variants support trainable per-splat attributes such as temperature or reflectance, facilitating joint radiometric and geometric optimization (106, 107, 108). Incorporating sensor-specific noise models and geometric priors during fitting increases metric fidelity and reduces modality-induced bias, an important consideration when continuous representations are used for analytic tasks that require metric reliability, such as thermal anomaly localization and construction monitoring (109, 87).

The differentiable nature of continuous renderers unlocks end-to-end training regimes where photometric, radiometric and geometric losses are optimized jointly. Such joint objectives permit simultaneous refinement of extrinsic calibrations, per-sample attribute integration and representation parameters, yielding unified models amenable to downstream inference. Yet this potential raises research challenges: multimodal sampling strategies must respect heterogeneous sampling densities and dynamic ranges, loss formulations must balance modality-specific errors, and optimization must avoid trivial solutions that overfit to a dominant modality. Recent studies explore mixed parameterizations that combine implicit fields with explicit Gaussian primitives to take advantage of continuous priors while preserving fast projectable primitives for rendering and compatibility with classical geospatial formats (110, 111). Streaming-friendly encodings, incremental fitting and sensor-aware regularization are active areas of research enabling continuous models to operate over repeated UAV passes and to support long-term monitoring tasks (103, 112).

From an application standpoint, continuous models have demonstrated strengths in high-fidelity visualization and in enabling differentiable analysis pipelines, but empirical evidence on their use for robust feature clustering in noisy, large-scale UAV-collected urban datasets is still nascent. Existing benchmarks and methodological reports show promising improvements in rendering quality and interactive visualization, but systematic studies on how continuous representations influence segmentation performance, thermal anomaly detection sensitivity and metric accuracy remain an open research direction (105, 113, 114).

The literature therefore suggests a twofold research agenda: extend continuous models to carry reliable multispectral attributes with uncertainty quantification, and evaluate how differentiable fitting and joint optimization affect downstream analytic tasks under realistic sensing conditions.

### 3.3 Unsupervised Feature Clustering and 3D Representation Learning

Feature clustering of large urban point clouds is challenged by class imbalance, clutter, occlusion and highly variable sampling density. Pioneering architectures like PointNet and PointNet++ introduced permutation-invariant handling of point sets and hierarchical local feature extraction, establishing a foundation for subsequent graph- and kernel-based innovations (11, 76). DGCNN, KPConv and RandLA-Net improved local geometric discrimination and computational scalability, enabling segmentation pipelines capable of operating on city-scale datasets with acceptable resource footprints (13, 115, 116). More recent transformer-based encoders and sparse-convolution hybrids extend receptive fields and capture long-range dependencies necessary for coherent city-scale contextual reasoning (117, 118, 119).

Recent work by Bultmann et al. (120) introduced a real-time integration framework for UAVs, integrating LiDAR, RGB, and thermal modalities with label propagation for cross-domain adaptation. Their system performs onboard supervised segmentation and inference, demonstrating the feasibility of multimodal data integration in constrained environments. In contrast, our work explores unsupervised feature clustering in multimodal point clouds, aiming for scalable and generalizable analysis under operational constraints, without reliance on annotated datasets or grouping priors.

The shortage of labeled data in many urban sensing scenarios has catalyzed a transition toward self-supervised and contrastive learning paradigms. Masked autoencoding for point clouds, patch-based contrastive objectives and geometry-aware consistency losses have been shown to produce transferable embeddings that generalize across scenes and sensor configurations, thereby reducing reliance on expensive manual annotation (14, 121, 122). Self-supervised pretraining followed by light-weight downstream adaptation attains competitive segmentation performance while dramatically reducing labeled-data requirements; large-scale pretraining on diverse urban datasets further improves consistency to noise and occlusion (123, 124).

Integrating radiometric attributes into pretraining and representation learning is crucial for leveraging thermal and multispectral signals. Methods that fuse geometric features with per-point radiometry during representation learning show improved class separability for thermally salient targets and increased sensitivity to material-dependent

anomalies (63, 125). Effective multimodal pretraining must carefully manage modality-specific augmentations and sampling schemes to preserve physical consistency; density-aware augmentations, balanced sampling across urban typologies and density-preserving downsampling are practical measures that reduce sampling bias and improve transferability (89, 88). Advances in ensemble encoder designs that combine global masked reconstruction with local graph-based feature extractors produce richer latent spaces that better support unsupervised clustering and panoptic grouping in heterogeneous outdoor scenes (126, 127).

Emerging research explores hybrid training regimes combining unsupervised pretraining with few-shot or semi-supervised fine-tuning to propagate sparse labels across large point clouds, effectively leveraging limited expert annotations to bootstrap large-scale segmentation tasks (128, 129). Additionally, temporal consistency and multi-pass fusion enable self-training loops where confident predictions from one pass provide pseudo-labels for subsequent refinement, a strategy particularly relevant for repeated UAV surveys and progressive map building (103, 114). Despite these advances, large-scale evaluations of unsupervised and semi-supervised protocols on realistic multimodal UAV benchmarks remain limited, and more extensive benchmarks are required to quantify generalization across sensors, seasons and acquisition conditions.

### 3.4 Recent Advances in 3D Reconstruction and Semantic Modeling

The field of three-dimensional reconstruction and semantic modeling has evolved rapidly over the past decade, driven by breakthroughs in deep learning, geometric attention mechanisms, and neural rendering. These advances have enabled systems to achieve higher fidelity, consistency, and generalization across diverse environments, particularly in urban and unstructured scenarios. The present work is situated within this evolving landscape and draws upon several foundational and emerging contributions from the literature.

One of the most significant evolutions in point cloud processing is the emergence of PointNeXt (130), a scalable and modular architecture that revisits PointNet++ with improved training strategies, residual connections, and hierarchical feature extraction. It achieves state-of-the-art performance in classification and segmentation tasks across benchmarks such as ScanObjectNN and ModelNet40. Its design supports multi-scale processing and efficient GPU acceleration, making it suitable for both coarse and fine-grained semantic inference. This architecture offers a promising alternative for unsupervised feature clustering, particularly in dense urban scenes with complex geometry and multimodal data. Complementing segmentation, the task of point cloud registration has seen notable progress with the introduction of GeoTransformer (131). This model incorporates geometry-guided attention by encoding pairwise distances and triplet-wise angles into transformer layers, enhancing correspondence accuracy in low-overlap and structurally ambiguous scenes. It

eliminates the need for traditional Random Sample Consensus (RANSAC) (132) post-processing by achieving high inlier ratios and registration recall, especially on benchmark datasets such as 3DLoMatch (133) and KITTI (Karlsruhe Institute of Technology and Toyota Technological Institute) (134). Its consistency and simplicity make it ideal for multimodal integration tasks, where spatial consistency between RGB, thermal, and LiDAR modalities is critical.

In the domain of direct 3D reconstruction, the Visual Geometry Grounded Transformer (VGGT) (38) represents a paradigm shift. VGGT infers depth maps, camera parameters, and dense point clouds directly from single or multi-view images using a feed-forward transformer architecture. It bypasses traditional optimization and triangulation steps, achieving sub-second inference times and enabling real-time reconstruction. VGGT has demonstrated superior performance in multi-view depth estimation, 3D point tracking, and novel view synthesis, and its pretrained backbone has been successfully transferred to downstream tasks such as non-rigid reconstruction and feature clustering. Although licensing constraints may limit its use in sensitive applications, VGGT remains a strategic reference for centralized reconstruction modules in hybrid systems.

Neural rendering has also advanced significantly with the development of Instant Neural Graphics Primitives (Instant-NGP) (135). This framework uses multiresolution hash encoding to accelerate training and rendering of neural graphics primitives, including Neural Radiance Fields (NeRFs) and signed distance functions (SDFs). It supports real-time reconstruction from image sequences and integrates seamlessly with pipelines like COLMAP (Structure-from-Motion and Multi-View Stereo) (136) for camera pose estimation. Its CUDA (Compute Unified Device Architecture) implementation enables deployment on RTX (NVIDIA Ray Tracing)-equipped systems, making it suitable for centralized control units in robotic swarms. Instant-NGP complements lightweight modules by providing high-fidelity refinement of spatial models, particularly in scenarios requiring photorealistic rendering or semantic overlays.

Simultaneous Localization and Mapping (SLAM) remains a foundational component of autonomous navigation and spatial understanding. Recent studies (137, 138) have explored SLAM implementations tailored for UAVs and resource-constrained platforms, addressing challenges such as localization drift, feature scarcity, and loop closure. Techniques like EKF-SLAM (Extended Kalman Filter for SLAM) (139), ORB-SLAM (Oriented FAST and Rotated BRIEF) (140), and Cartographer (141) offer real-time mapping and pose estimation using visual, LiDAR, and inertial data. Innovations include visual-reference coupling, sparse matrix optimization, and neural denoising modules for embedded deployment. These methods are highly compatible with onboard perception goals, enabling incremental mapping and spatial anchoring in dynamic environments. Their integration opens new possibilities for autonomous navigation, swarm coordination, and multitemporal

reconstruction.

To support the implementation and testing of these advanced methods, several open-source frameworks have emerged as essential tools. Open3D (68) provides a comprehensive library for point cloud processing, visualization, and registration, with support for ICP, voxel grids, and RGBD integration. Torch Points3D (142) offers modular implementations of deep learning architectures for 3D data, including PointNet++, DGCNN, KPConv, and RandLA-Net. Instant-NGP and Kaolin (143) extend support for neural rendering and differentiable graphics, facilitating experimentation with NeRFs and mesh-based models. These frameworks enhance the reproducibility and extensibility of research in clustering modeling and reconstruction, and serve as foundational components for the continued evolution of multimodal, unsupervised, and real-time spatial intelligence.

In summary, the recent literature provides a rich and diverse foundation for the development of systems capable of operating across heterogeneous environments. The integration of geometry-aware attention, scalable segmentation, neural rendering, and SLAM techniques not only informs current implementations but also guides future extensions toward more autonomous, adaptive, and semantically coherent spatial modeling. These advances reflect a broader shift in the field toward architectures that are not only accurate and efficient, but also deployable in real-world robotic platforms.

### 3.5 Gaps, Practical Constraints and Reproducibility

A critical synthesis of the reviewed literature reveals persistent gaps that limit operational adoption. Integrated end-to-end systems that preserve semantic and metric fidelity across RGB, thermal and LiDAR modalities are still scarce, with most studies addressing only subsets of the full multimodal stack or relying on dense supervision to reach high accuracy (86, 144). The scarcity of public, large-scale multimodal UAVs datasets with calibrated thermal and LiDAR ground truth hinders robust benchmarking and transfer learning studies; where datasets exist they often reflect limited urban typologies or controlled conditions (98, 99, 29). Reproducibility is further challenged by dependence on proprietary toolchains, heavy computational requirements for continuous models and heterogeneous formats for multispectral attribute storage (68, 101, 112).

From a methodological perspective, combining continuous neural renderers with unsupervised feature clustering introduces optimization and evaluation complexities. Joint objectives must balance geometric, photometric and radiometric losses while accounting for modality-specific noise and uncertainty; without careful loss weighting and uncertainty modeling, optimization can overfit dominant modalities or produce artifacts that degrade downstream clustering quality (87, 145). Operational constraints further impose trade-offs: real-time tasks require compact, streaming-compatible encodings and incremental update

schemes, whereas high-fidelity forensic analyses permit longer processing times but demand strict metric accuracy and traceable provenance (102, 103).

Addressing these gaps requires coordinated efforts on several fronts. Curating large, diverse and well-calibrated multimodal UAVs datasets with standardized metadata and grouping labels is essential to support fair comparisons. Developing benchmark protocols and challenge suites that evaluate combined reconstruction, radiometric fidelity and grouping inference under variable environmental conditions will clarify strengths and limitations of proposed methods. Engineering reference implementations and open-source pipelines that demonstrate staged deployment patterns will facilitate transitions from research prototypes to operational systems and allow practitioners to compare end-to-end trade-offs in compute, latency and accuracy (28, 114, 101).

### 3.6 Comparative Synthesis and Rationale for GaussianFusion\_AI

The reviewed literature exposes clear trade-offs across methodological choices. Supervised voxelization and 3D-convolutional strategies achieve high segmentation accuracy on labeled datasets but induce geometric coarsening and are vulnerable to label scarcity (146). Contrastive spectral approaches enhance discrimination for thermal or material-specific classes yet frequently omit full 3D reconstruction and LiDAR alignment, reducing metric reliability for spatial analytics (147, 148). Continuous renderers based on Gaussian formulations enable compact, projectable representations and interactive visualization, but integrated demonstrations combining explicit multimodal integration with unsupervised feature clustering remain nascent (2, 108). A hybrid design that couples continuous multimodal encoding, an ensemble of unsupervised encoders, and modular engineering practices offers a pragmatic path to reconcile representation quality, grouping inference, and operational reproducibility.

`GaussianFusion_AI` is motivated by these observations: continuous Gaussian-based splatting produces projectable, attribute-rich primitives that remain compatible with point-based processing and GIS workflows; ensemble encoder designs combine complementary global and local cues to strengthen unsupervised feature clustering; and a modular pipeline enables staged calibration, data integration, and inference that respect operational constraints. The architecture demonstrates that joint optimization of geometric registration, radiometric alignment, and unsupervised representation learning yields reproducible, semantically enriched 3D products suitable for operational diagnostics such as thermal-anomaly detection and urban-morphology monitoring.

### 3.7 Contextualization and Selected Literature

To contextualize this proposal within the state of the art, we selected a representative set of recent works that reflect key directions in multimodal data integration, 3D segmentation, and differentiable rendering. Selection criteria prioritized diversity in modality integration (RGB, thermal, LiDAR), methodological approaches (supervised learning, heuristic detection, contrastive learning, Transformer architectures), and rendering strategies (voxelization, rasterization, splatting). These works were chosen for their relevance to geospatial analysis and their influence on current research trends. While some of them explicitly involve UAV-based acquisition, others focus on general-purpose 3D datasets or indoor/outdoor benchmarks; the common thread is their methodological contribution to multimodal data integration and representation learning, rather than the acquisition platform itself.

Each of these studies offers a distinct perspective on the challenges of semantic modeling, data integration, and rendering efficiency. At the same time, they present limitations that motivate further investigation. For instance, voxelization-based methods often suffer from high memory consumption and reduced scalability in large urban scenes; rasterization approaches can lose geometric fidelity when projecting multimodal attributes; splatting techniques, though efficient, are still constrained by dataset availability and sensor synchronization issues. Transformer-based segmentation models demonstrate strong performance but remain dependent on large annotated datasets, which are scarce in multimodal UAV contexts. Contrastive learning frameworks improve generalization but struggle with heterogeneous sampling densities and temporal drift across modalities.

By highlighting these limitations, the literature review establishes the need for architectures that balance efficiency, consistency, and multimodal adaptability. The subsequent chapter will present the proposed framework in detail, showing how it addresses the gaps identified here.

Representative studies also illustrate persistent limitations that motivate the present proposal. Zhou et al. (2023) presented a supervised RGB–LiDAR fusion approach using voxelization and 3D convolutional networks for segmentation; while semantically effective, the method depends on manual labels, excludes thermal data, and lacks continuous rendering, with voxelization introducing computational overhead and geometric detail loss. The reliance on discrete voxel grids hinders scalability and adaptability to large-scale outdoor UAV scenarios.

Ma et al. (2024) explored RGB–thermal contrastive learning to enhance feature discrimination across modalities. However, their framework omits 3D reconstruction, LiDAR integration and field validation, limiting applicability to real-world geospatial tasks. The absence of spatial depth cues and volumetric modeling restricts utility in topographic

or structural mapping, where multimodal depth integration is essential.

Chen et al. (2024) proposed a Transformer-based 2.5D segmentation approach operating on rasterized depth maps. While effective in structured indoor scenes, rasterization discards fine-grained geometric detail and precludes continuous rendering. The model’s dependence on depth maps constrains generalization to unstructured outdoor environments and UAV-acquired point clouds with thermal gradients.

Gholipour et al. (2024) combined thermal and LiDAR modalities for urban energy mapping using static heuristics to detect thermal hotspots. Although the thermal–geometric integration is promising, the absence of learning-based segmentation and differentiable rendering limits adaptability and precision; the approach lacks semantic generalization and cannot dynamically adjust to varying urban morphologies or material properties.

Kerbl et al. (2023) introduced 3D Gaussian-based splatting with high rendering efficiency and realism, marking a breakthrough in continuous volumetric rendering. Original demonstrations focused on SfM-derived reconstructions and did not integrate semantic attributes or multimodal integration. The lack of modality-aware learning and semantic segmentation restricts direct application in geospatial analytics and autonomous navigation.

Table 1 summarizes a technical comparison between these works and the present proposal.

Table 1 – Comparison between recent works and this proposal

Work	RGB	Thermal	LiDAR	3D Segmentation	Continuous Rendering
Zhou et al. (2023)	Yes	No	Yes	Supervised	No
Ma et al. (2024)	Yes	Yes	No	Supervised	No
Chen et al. (2024)	Yes	No	No	Supervised	No
Gholipour et al. (2024)	No	Yes	Yes	Heuristic	No
<b>This work</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Unsupervised</b>	<b>Yes</b>

The comparative analysis highlights that `GaussianFusion_AI` uniquely integrates UAV-acquired RGB–thermal–LiDAR datasets, continuous Gaussian-based splatting, unsupervised 3D feature clustering, and a modular open-source implementation. Unlike prior works, this approach supports continuous rendering and semantic inference across heterogeneous modalities without reliance on manual annotations or rasterization. The integration of thermal gradients, geometric depth and visual texture within a differentiable rendering pipeline enables scalable deployment in outdoor, dynamic environments. The unsupervised learning paradigm enhances generalization and reduces annotation costs, positioning the system for large-scale geospatial applications such as infrastructure monitoring, energy auditing and autonomous navigation.

## Conclusion

This chapter presented a consolidated review of the scientific foundations relevant to the `GaussianFusion_AI` architecture. By examining multimodal sensor data integration, continuous neural representations and unsupervised 3D inference, the review identified methodological gaps and operational barriers that the proposed system aims to address. In particular, the scarcity of large-scale multimodal datasets, the limitations of conventional registration techniques such as ICP, and the dependency of segmentation models on extensive manual annotations were highlighted as persistent challenges in the field.

The comparative analysis of related works demonstrated that, although significant progress has been achieved in supervised segmentation, neural rendering, and multimodal alignment, existing approaches often remain constrained by memory demands, annotation costs, or restricted modality integration. These limitations underscore the need for frameworks that are both scalable and adaptable to heterogeneous urban environments, capable of integrating thermal, geometric, and visual cues in a unified representation.

By synthesizing these insights, the chapter establishes a clear rationale for advancing toward architectures that combine efficiency, consistency, and semantic coherence. The next chapter builds directly on this foundation, detailing the methodological framework, implementation choices, and experimental protocols designed to validate the proposed system in operational urban scenarios. This transition marks the shift from theoretical grounding to practical realization, where the concepts reviewed here are transformed into an integrated solution for large-scale geospatial intelligence.

## 4 HYBRID ARCHITECTURE FOR FEATURE CLUSTERING OF UAV MULTIMODAL DATA

In recent years, the use of drones equipped with advanced sensors has changed how three-dimensional data are captured and analyzed. Combining optical images (RGB), thermal measurements and depth returns from LiDAR sensors has enabled new applications in urban modeling, heritage inspection, environmental monitoring and beyond. Building on this context, this research proposes a modular, integrated system, `GaussianFusion_AI`, designed to transform heterogeneous data into detailed, continuous and semantically informative 3D models.

The distinguishing feature of this work is the integration of sensor information from the outset. Instead of processing each modality separately and fusing results afterwards, the methodology adopts a native multimodal strategy in which all data are combined collaboratively. The resulting point cloud is converted into a continuous representation using Gaussian-based splatting, producing smooth visualizations that avoid the gaps typical of sparse clouds. On this continuous basis, neural networks such as PointNet++, DGCNN and Point-MAE are applied to perform feature clustering that partitions the scene into meaningful regions (façades, vegetation, sidewalks) without requiring manual labels.

This chapter details the full processing chain: platforms and sensors used; preprocessing and multimodal integration; unsupervised 3D feature clustering; and quantitative evaluation. The methodology was tested in three urban scenarios in Rio de Janeiro (CTEx, Outeiro da Glória and Quinta da Boa Vista), each posing specific challenges and contributing to validate the proposal’s consistency.

Beyond describing a technical system, the chapter explains how the combination of sensing, computer vision and data science yields rich, useful and efficient 3D models for real-world applications. The approach aims to interpret cities with greater depth, accuracy and intelligence.

### 4.1 Theoretical Foundations and Research Framing

The hybrid architecture `GaussianFusion_AI` emerges as a methodological response to persistent limitations in multimodal UAV-based analysis. While RGB, thermal and LiDAR sensors offer complementary perspectives (capturing texture, emissivity and geometric structure), their integration into coherent and semantically meaningful representations remains a challenge. Conventional pipelines often rely on late fusion strategies, supervised feature labeling and discrete rendering, which restrict scalability, semantic generalization

and adaptability to real-world variability (149). These limitations are particularly evident in urban environments, where occlusion, material heterogeneity and dynamic conditions demand more flexible and intelligent modeling approaches.

To verify that embeddings from PointNet++, EdgeConv and Point-MAE provide complementary rather than redundant information, pairwise correlation and mutual information metrics were computed. Each network emphasized distinct aspects of the data: PointNet++ captured multiscale local geometry, EdgeConv highlighted topological relations, and Point-MAE encoded contextual semantics. Ablation studies confirmed this complementarity, as removing one branch reduced clustering coherence and silhouette scores.

Thermal distortion was mitigated through radiometric calibration and temporal alignment. Raw thermal frames were corrected using manufacturer calibration curves and normalized against ambient baselines to reduce drift. Motion-aware interpolation ensured that thermal attributes were consistently mapped to geometric points, minimizing parallax and frame-rate mismatches between sensors.

Normalization of multimodal attributes (RGB, temperature, LiDAR intensity, covariance eigenvalues) was performed with modality-specific scaling. Each modality was normalized independently to preserve its physical meaning, and embeddings were projected into a shared latent space using learnable scaling factors. This strategy prevented heterogeneous units (e.g., degrees Celsius vs. reflectance intensity) from distorting geometric relationships, while allowing the networks to balance contributions from different modalities adaptively.

This research is motivated by the need for a unified framework capable of producing interpretable, continuous and unsupervised representations of complex scenes. The central hypothesis is that early-stage integration of multimodal UAV data, combined with latent-space feature clustering and continuous rendering, enables semantic modeling without manual annotation or rasterization. This hypothesis is tested through the implementation of a hybrid architecture that integrates sensing, learning and rendering in a single pipeline, designed to operate across varied urban morphologies and sensor configurations. The guiding research question is: *How can a hybrid architecture based on unsupervised representation learning and continuous rendering optimize the semantic interpretation of real environments through integration of RGB, thermal and LiDAR data acquired by UAVs?* This formulation reflects a conceptual shift from geometric reconstruction to semantic understanding, emphasizing the role of unsupervised modeling and continuous representations in extracting meaning from multimodal observations. It also aligns with emerging paradigms in geospatial intelligence, where the goal is not merely to visualize but to interpret spatial phenomena algorithmically (150).

The general objective of this work is to develop and validate a scalable architecture for semantic interpretation of UAV-acquired multimodal data. To achieve this, the methodology encompasses the calibration and registration of RGB, thermal and LiDAR point clouds into a unified spatial frame, the application of unsupervised feature clustering with PointNet++, DGCNN and Point-MAE to capture complementary geometric and contextual cues, the integration of Gaussian-based splatting for continuous rendering and multimodal attribute overlay, and the evaluation of performance using geometric and structural metrics such as RMSE, Chamfer Distance, Hausdorff Distance and silhouette index. Reproducibility is ensured through the publication of code, datasets and scripts in an open repository, fostering collaborative development and transparency.

Together, these elements contribute directly to the overarching goal of enabling robust, label-free semantic interpretation of multimodal UAV data in complex urban environments. By consolidating sensing, representation learning and continuous rendering, the proposed framework addresses persistent limitations in scalability, annotation dependency and modality integration, positioning itself as a methodological advance for large-scale geospatial applications such as infrastructure monitoring, energy auditing and autonomous navigation. These contributions inform the design of `GaussianFusion_AI`, which combines early integration, unsupervised feature clustering and continuous rendering to produce semantically expressive models. The architecture extends beyond 3D modeling to semantic interpretation, supporting applications in urban analytics, infrastructure monitoring and environmental diagnostics. Its modularity also enables future extensions, including new modalities (e.g., hyperspectral imaging), learning paradigms (e.g., contrastive or generative approaches) and deployment contexts (e.g., post-disaster assessment, cultural heritage preservation or smart city planning).

Moreover, the architecture responds to a growing demand for interpretable AI in geospatial domains. As highlighted by recent reviews (150, 149), the integration of explainable models with multimodal data remains a critical challenge. `GaussianFusion_AI` contributes to this discourse by offering a transparent pipeline, where each stage—from sensor integration to feature clustering and rendering—is documented, reproducible and open to scrutiny. The use of open-source libraries such as Open3D, PyTorch and NumPy ensures accessibility, while the repository [https://gitlab.com/tjmb\\_ime/GaussianFusion\\_AI](https://gitlab.com/tjmb_ime/GaussianFusion_AI) provides code, datasets and experimental scripts for community use, reinforcing the principles of Open Science.

This section frames the theoretical and methodological foundations of the proposed architecture, articulating its relevance, originality and potential impact. By bridging sensing, learning and rendering in a unified system, `GaussianFusion_AI` advances the state of the art in multimodal interpretation and lays the groundwork for future research in intelligent spatial modeling.

## 4.2 Overview of the Experimental Architecture

The proposed architecture, named `GaussianFusion_AI`, is designed to reconstruct urban environments from multimodal aerial data by integrating RGB imagery, thermal measurements and LiDAR scans into a unified 3D representation. The pipeline is modular and reconfigurable, allowing each stage to be independently refined while maintaining compatibility with the overall flow. Its design emphasizes reproducibility, clarity of processing stages and adaptability to diverse urban scenarios.

Figure 12 presents the consolidated conceptual flow of the architecture. Each block corresponds to a distinct module, with names aligned to the subsections that describe them, forming a coherent sequence from raw data acquisition to rendering.

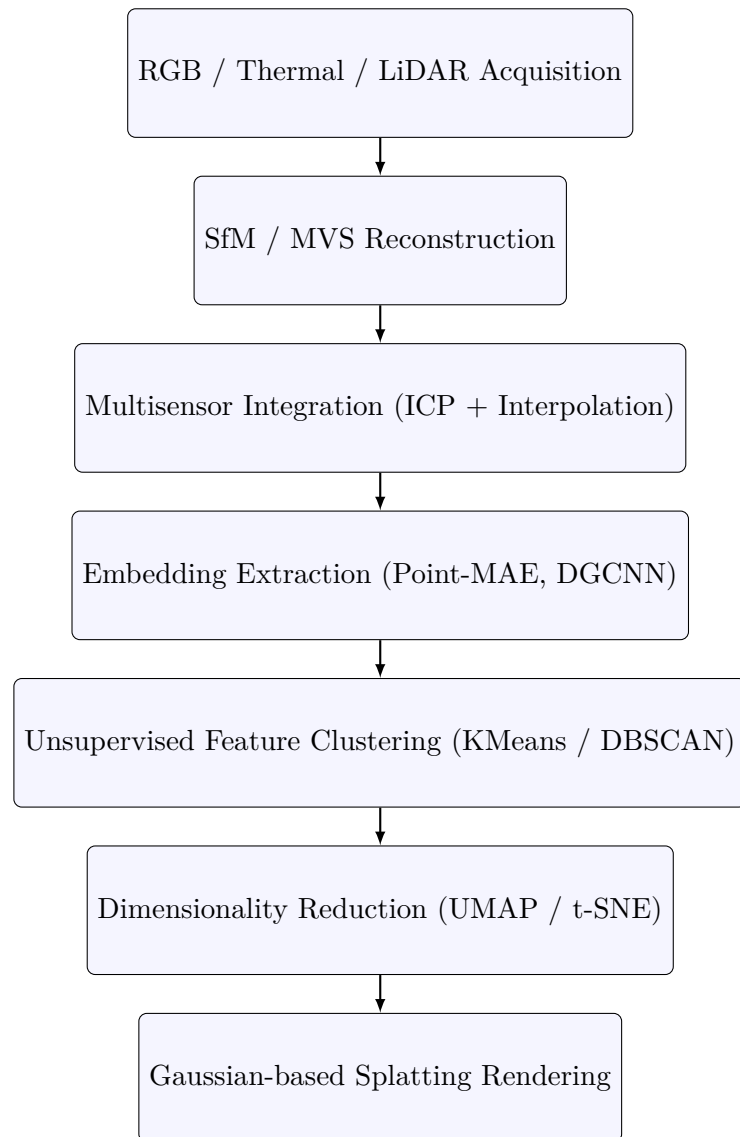


Figure 12 – Conceptual pipeline of the `GaussianFusion_AI` architecture. Each module processes multimodal inputs toward an unsupervised feature-clustered and continuously rendered 3D scene using Gaussian-based splatting.

The pipeline begins with multimodal data acquisition using drones equipped with RGB cameras, thermal imagers and LiDAR sensors. These platforms collect spatially and radiometrically diverse datasets over urban areas with varying architectural complexity and material heterogeneity.

The reconstruction stage applies Structure-from-Motion and Multi-View Stereo to RGB and thermal imagery, generating dense point clouds. LiDAR data is preprocessed and georeferenced. All modalities are registered into a common Euclidean frame using rigid alignment techniques such as ICP, ensuring spatial consistency and enabling integration.

Multisensor integration combines the registered point clouds into a unified structure enriched with geometric, photometric and thermal attributes. Interpolation techniques fill radiometric gaps, and the resulting cloud is exported in `.ply` format to preserve per-point metadata and facilitate downstream processing.

Embedding extraction is performed using neural architectures such as Point-MAE and DGCNN, which learn latent representations that encode local geometry, topological relations and contextual information. These embeddings are then clustered using unsupervised algorithms. K-Means was adopted as a baseline due to its simplicity and efficiency, while DBSCAN complemented the analysis by capturing clusters of arbitrary shapes and densities. Hyperparameters for these algorithms, as well as for dimensionality reduction methods (UMAP, t-SNE), were tuned empirically using silhouette scores and stability measures to ensure meaningful partitions of the latent space.

To visualize and validate the embeddings, dimensionality reduction techniques such as UMAP and t-SNE are applied. These projections expose the structure of the learned representations and support the interpretation of clustering results.

Finally, the integrated cloud is rendered using Gaussian-based splatting. Each point is modeled as an anisotropic Gaussian with spatial support, opacity and multimodal attributes. The splats are projected into the image plane and composited using differentiable rasterization, producing smooth, continuous rendering suitable for inspection, analysis and integration with learning pipelines.

In addition to the conceptual pipeline, a computational-level flowchart was designed to illustrate how data move between modules, from raw sensor files to embeddings, clusters, projections and rendered outputs. This dual perspective clarifies both the methodological design and the practical implementation of the architecture.

### 4.3 Preprocessing and Multimodal Structuring

Preprocessing and multimodal structuring are crucial to ensure that each sensor contributes coherently to the final model. Before converting the integrated clouds into

a continuous representation, optical, thermal and LiDAR measurements must share a common geometric reference and their attributes must be calibrated and consolidated. This stage is divided into three main tasks: photogrammetric reconstruction for RGB and thermal images, geometric processing and filtering of LiDAR clouds, and geometric integration with attribute interpolation.

### 4.3.1 Photogrammetric reconstruction of RGB and thermal data

From RGB and thermal image sets we run a four-stage reconstruction pipeline: feature detection, sparse SfM, depth densification MVS, and association of thermal channels to reconstructed points.

During feature matching, images are processed with descriptors such as SIFT (45) or ORB (47). Corresponding points must satisfy the epipolar constraint. The homogeneous pixel coordinates  $\mathbf{u}_i$ ,  $\mathbf{u}_j$  and the fundamental matrix  $\mathbf{F}$  satisfy:

$$\mathbf{u}_j^\top \mathbf{F} \mathbf{u}_i = 0 \quad (4.1)$$

SfM jointly estimates camera poses and sparse 3D structure by minimizing reprojection error:

$$\min_{\{\mathbf{R}_i, \mathbf{t}_i\}, \{X_j\}} \sum_{i=1}^N \sum_{j=1}^M \|\mathbf{u}_{ij} - \pi(\mathbf{R}_i X_j + \mathbf{t}_i)\|^2 \quad (4.2)$$

Solved with Levenberg–Marquardt using libraries such as OpenMVG or COLMAP. After sparse reconstruction, multiview stereo produces a dense point cloud  $\mathcal{P}_{\text{RGB}}$ .

Thermal channel incorporation projects each reconstructed point  $X_j$  into thermal images:

$$\mathbf{u}_{ij} = \pi(\mathbf{R}_i X_j + \mathbf{t}_i). \quad (4.3)$$

From the matched thermal pixel we interpolate brightness temperature  $T_j$  and perform min–max normalization:

$$\tau_j = \frac{T_j - T_{\min}}{T_{\max} - T_{\min}} \quad (4.4)$$

So  $\tau_j \in [0, 1]$ . The resulting point attributes are  $(x_j, y_j, z_j, r_j, g_j, b_j, \tau_j)$ , forming  $\mathcal{P}_{\text{opt+therm}}$ . Here,  $(x_j, y_j, z_j)$  denote the 3D spatial coordinates of point  $j$ ,  $(r_j, g_j, b_j)$  are the radiometric values from the RGB image, and  $\tau_j$  is the normalized thermal attribute derived from brightness temperature. Together, these attributes encode geometry, color and thermal information in a unified multimodal point cloud.

### 4.3.2 Geometric processing of the LiDAR cloud

LiDAR returns are initially expressed in spherical coordinates  $(r, \theta, \phi)$ , where  $r$  is the radial distance from the sensor,  $\theta$  the azimuth angle and  $\phi$  the elevation angle. These are converted to Cartesian coordinates by:

$$x = r \sin \phi \cos \theta, \quad y = r \sin \phi \sin \theta, \quad z = r \cos \phi. \quad (4.5)$$

Statistical filtering removes outliers: a point is classified as noisy if the mean distance to its  $k$  nearest neighbors exceeds  $\mu + 3\sigma$ , where  $\mu$  and  $\sigma$  denote the mean and standard deviation of neighborhood distances. This criterion follows the Statistical Outlier Removal method widely adopted in point cloud processing (151). When multiple returns are available, they are preserved to capture surface complexity. Local surface normals are estimated via neighborhood covariance. For a point  $p$  with neighbors  $\{q_i\}$ , the covariance matrix is:

$$\mathbf{C}_p = \frac{1}{k} \sum_{i=1}^k (q_i - p)(q_i - p)^\top \quad (4.6)$$

The eigenvector associated with the smallest eigenvalue of  $\mathbf{C}_p$  defines the local normal  $\mathbf{n}_p$ , representing the orientation of the surface at point  $p$ .

### 4.3.3 Multimodal Integration and Cross Interpolation

After registration, two consolidation steps are performed: rigid registration with ICP and attribute interpolation using distance-weighted neighborhoods.

Rigid registration estimates the rigid transformation  $(\mathbf{R}, \mathbf{t}) \in SE(3)$ , where  $\mathbf{R}$  is a rotation matrix and  $\mathbf{t}$  a translation vector, aligning the LiDAR cloud  $\mathcal{P}_{\text{LiDAR}}$  to the optical–thermal cloud  $\mathcal{P}_{\text{opt+therm}}$  by minimizing:

$$\min_{\mathbf{R}, \mathbf{t}} \sum_{p \in \mathcal{P}_{\text{LiDAR}}} \min_{q \in \mathcal{P}_{\text{opt+therm}}} \|\mathbf{R}p + \mathbf{t} - q\|^2. \quad (4.7)$$

Here,  $p$  and  $q$  denote 3D points in the LiDAR and optical–thermal clouds, respectively. Implementation uses Open3D’s `registration_icp()` with tuned parameters.

For multisensor attribute interpolation, for each LiDAR point  $p$  we find its  $k$  nearest neighbors  $\{q_i\}$  in the optical/thermal cloud and compute Gaussian-weighted interpolation:

$$\tilde{\mathbf{c}}(p) = \frac{\sum_{i=1}^k \exp\left(-\frac{\|p-q_i\|^2}{\sigma^2}\right) \cdot \mathbf{c}(q_i)}{\sum_{i=1}^k \exp\left(-\frac{\|p-q_i\|^2}{\sigma^2}\right)}. \quad (4.8)$$

In this expression,  $\mathbf{c}(q_i)$  represents the attribute vector of neighbor  $q_i$  (color and thermal values),  $\|p - q_i\|$  is the Euclidean distance between points, and  $\sigma$  controls the Gaussian kernel width.

The integrated point  $p_i$  is then constructed as:

$$p_i = [x_i, y_i, z_i, r_i, g_i, b_i, \tau_i, n_x^i, n_y^i, n_z^i, I_i^{\text{LiDAR}}]. \quad (4.9)$$

Here,  $(x_i, y_i, z_i)$  are the Cartesian coordinates,  $(r_i, g_i, b_i)$  the RGB values,  $\tau_i$  the normalized thermal attribute,  $(n_x^i, n_y^i, n_z^i)$  the components of the estimated surface normal, and  $I_i^{\text{LiDAR}}$  the LiDAR intensity return. The integrated cloud is exported as `.ply` or `.npz` for compatibility with Open3D and PyTorch Geometric.

#### 4.3.4 Continuous modeling with Gaussian-based splatting

To overcome sparsity and irregular sampling, each integrated point  $p_i$  is represented by an anisotropic 3D Gaussian. The continuous contribution of a single element is:

$$\mathcal{G}_i(x) = \alpha_i \cdot \exp\left(-\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)\right), \quad (4.10)$$

with  $x \in \mathbb{R}^3$ . Here  $\mu_i$  is the point center,  $\Sigma_i$  is a local covariance estimated by PCA on the neighborhood, encoding anisotropy and orientation, and  $\alpha_i \in [0, 1]$  is an opacity weight derived from local density and sensor confidence. Color and normalized thermal attributes are attached to each Gaussian.

The practical implementation used in this work intentionally differs from fully differentiable Gaussian-based splatting pipelines. Instead of performing gradient-based optimization of Gaussian parameters through differentiable rasterization, the current module initializes splats directly from the integrated point cloud and textures them by projective mapping from RGB images. Rendering of the splats is performed from a chosen viewpoint using a forward compositing step with a soft z-ordering to approximate occlusion. The workflow is summarized in Figure 13.

It is important to note that the pipeline is not strictly end-to-end. While most stages are automated, human intervention occurs in the choice of hyperparameters (e.g., neighborhood size  $k$ , kernel width  $\sigma_p$ ), in the tuning of ICP parameters, and in the manual selection of viewpoints for rendering. These adjustments ensure stability and visual interpretability but also highlight the distinction from fully differentiable pipelines. The per-pixel compositing is expressed as:

$$I(u, v) = \frac{\sum_i w_i(u, v) C_i}{\sum_i w_i(u, v)}, \quad w_i(u, v) = \alpha_i \cdot \exp\left(-\frac{(d_i(u, v))^2}{2\sigma_p^2}\right), \quad (4.11)$$

where  $I(u, v)$  is the rendered pixel color at image coordinates  $(u, v)$ ,  $C_i$  is the RGB color attached to Gaussian  $i$ ,  $d_i(u, v)$  is the projected 2D distance from the splat center to  $(u, v)$ , and  $\sigma_p$  controls the projected spread.

Because the module does not perform multi-view gradient refinement of  $\mu_i, \Sigma_i, \alpha_i$ , the result is best described as a Gaussian-based splatting visualization pipeline rather than as an instantiation of canonical differentiable Gaussian-based splatting. The distinction is important for reproducibility and for setting expectations about multi-view consistency: the current implementation produces single-view continuous renderings suitable for inspection and overlay, but it may present occlusion artifacts in dense or overlapping regions that a full differentiable refinement would mitigate. Appendix B provides implementation details, parameter values and comparative notes. The workflow stages—splat initialization, texturing and viewpoint compositing—are illustrated in Figure 13.

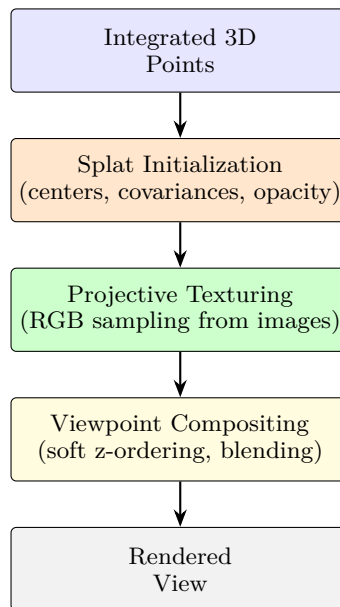


Figure 13 – Gaussian-based splatting module used in `GaussianFusion_AI`. The module initializes splats from the integrated cloud, applies image-based texturing, and composes a viewpoint-dependent rendering using soft blending.

## 4.4 Feature Extraction and Clustering

After building the continuous Gaussian-based representation and enriching each element with attributes, the pipeline extracts latent representations and performs clustering to partition the scene into coherent regions. References to supervised semantic labeling are intentionally avoided; the pipeline focuses on unsupervised latent representations, feature clustering and region discovery.

#### 4.4.1 Overall flow of feature extraction and clustering

Figure 14 summarizes the unsupervised feature-clustering pipeline: data input, feature extraction, latent clustering and clustered point-cloud output.

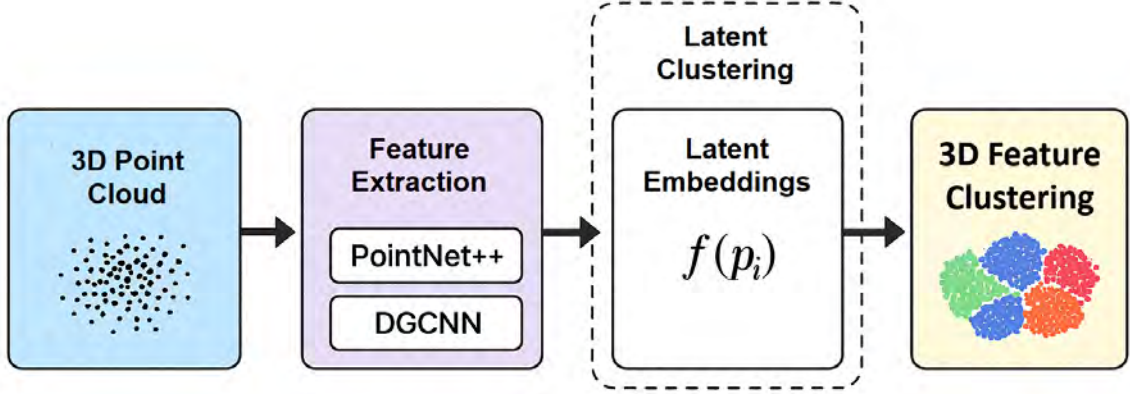


Figure 14 – Pipeline for unsupervised feature clustering based on latent embeddings; outputs are clustered point clouds used for downstream analysis and continuous rendering with Gaussian-based splatting.

The flow consists of input vector construction, embedding extraction by neural encoders, dimensionality reduction for inspection, and clustering for region assignment. Input is the multimodal cloud  $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^N$  with attribute vectors:

$$\mathbf{p}_i = [x_i, y_i, z_i, r_i, g_i, b_i, \tau_i, n_x^i, n_y^i, n_z^i, I_i]. \quad (4.12)$$

Where  $(x_i, y_i, z_i)$  are positions,  $(r_i, g_i, b_i)$  color values,  $\tau_i$  normalized temperature,  $(n_x^i, n_y^i, n_z^i)$  surface normals, and  $I_i$  LiDAR intensity. Feature extraction is performed by PointNet++, DGCNN and Point-MAE, producing embeddings  $f(\mathbf{p}_i) \in \mathbb{R}^{128}$ . Clustering was conducted primarily with DBSCAN, selected for its consistency to non-spherical clusters, while k-Means and HDBSCAN were used for comparative evaluation. This yields labels  $l_i$  assigned to points, producing  $\{(\mathbf{p}_i, l_i)\}$  for visualization and quantitative assessment.

#### 4.4.2 Input vector and preprocessing

Each Gaussian  $\mathcal{G}_i$  is converted into a descriptive vector:

$$v_i = [\mu_i; \text{eig}_1(\Sigma_i), \text{eig}_2(\Sigma_i), \text{eig}_3(\Sigma_i); \alpha_i; r_i, g_i, b_i; t_i; n_i], \quad (4.13)$$

Where eigenvalues capture local spread,  $\alpha_i$  is opacity,  $t_i$  is normalized thermal value and  $n_i$  the unit normal. All components are normalized prior to encoder input.

### 4.4.3 Neural networks for embedding extraction

Normalized vectors  $v_i$  are processed by a hybrid architecture composed of three specialized networks:

- **PointNet++**: hierarchical multiscale set abstraction (Ball Query / k-NN) combined with mini-PointNets to extract local features.
- **DGCNN**: dynamic graph convolution using EdgeConv to capture topological relations via local differential information.
- **Point-MAE**: masked autoencoder with Transformer-based encoder–decoder that reconstructs masked tokens, encouraging robust contextual embeddings.

Outputs of the three networks are concatenated and passed to an MLP:

$$f(p_i) = \text{MLP} \left( \text{Concat} \left[ \text{PointNet++}(v_i), \text{EdgeConv}(v_i), \text{MAE}(v_i) \right] \right), \quad (4.14)$$

producing a 128-dimensional latent vector  $f(p_i)$ .

### 4.4.4 Clustering in the latent space

Embeddings  $\{f(p_i)\}$  are clustered using k-Means and density-based methods. In practice, each network produces a latent vector:  $f_{\text{PN++}}(p_i)$ ,  $f_{\text{DGCNN}}(p_i)$ , and  $f_{\text{MAE}}(p_i)$ . These are concatenated into a single representation

$$f(p_i) = \left[ f_{\text{PN++}}(p_i) \parallel f_{\text{DGCNN}}(p_i) \parallel f_{\text{MAE}}(p_i) \right], \quad (4.15)$$

where  $\parallel$  denotes vector concatenation. No weighting scheme or fine-tuning was applied; the integration is purely additive, intended to preserve complementary geometric and contextual cues from each encoder. Future work may explore weighted contributions or joint training to validate theoretical optimality.

k-Means minimizes within-cluster squared distances:

$$\min_{C_1, \dots, C_K} \sum_{i=1}^K \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2, \quad (4.16)$$

with  $K$  chosen by elbow and silhouette analysis. DBSCAN groups by local density: a point  $x$  is a core if

$$|\{y \in \mathcal{P} : \|x - y\| \leq \epsilon\}| \geq \text{MinPts}, \quad (4.17)$$

allowing detection of arbitrary-shaped clusters and noise handling. HDBSCAN is used when hierarchical density analysis is needed.

Cluster labels are written to the `class` field in exported `.ply` files for colored visualization and further geometric validation.

#### 4.4.5 Visualization and qualitative inspection

Splat Gaussians are rendered with colors mapped to cluster labels, enabling inspection of separation between planar, vertical and organic regions. Opacity and thermal overlays help confirm correlations between physical attributes and discovered regions. UMAP is applied to project embeddings  $f(p_i) \in \mathbb{R}^{128}$  into  $\mathbb{R}^3$  for exploratory analysis; UMAP loss is expressed as:

$$\mathcal{L}_{\text{UMAP}} = \sum_{(i,j)} w_{ij} \|u_i - u_j\|^2, \quad (4.18)$$

where  $u_i$  is the projected 3D coordinate and  $w_{ij}$  weights neighborhood connectivity. Spanning Neighborhood Tree (SNT) structures are constructed over the projected points to reveal continuous relations and transitions.

## 4.5 Quantitative Evaluation of Reconstruction and Clustering

To guarantee comparability and reproducibility across methods, the evaluation protocol implements metrics that jointly assess geometric fidelity of 3D reconstructions and structural quality of clustering outputs. Metric choice follows established practice in point-set and clustering validation (152, 153, 122).

### 4.5.1 Reconstruction Evaluation

The geometric fidelity of reconstructed point clouds is quantitatively assessed by comparing each reconstruction  $P = \{p_i\}_{i=1}^N$  to a reference cloud  $Q$ . Three complementary metrics are employed to capture distinct aspects of reconstruction quality: global error, average proximity and worst-case deviation.

The Root Mean Square Error (RMSE) measures global geometric accuracy and is defined as:

$$\text{RMSE}(P, Q) = \sqrt{\frac{1}{N} \sum_{i=1}^N \|p_i - q_i\|^2} \quad (4.19)$$

Where  $q_i$  denotes the corresponding reference point for  $p_i$ . RMSE is sensitive to systematic misalignments and large-scale bias, making it suitable for evaluating overall registration and scale consistency.

The Chamfer Distance (CD) evaluates mutual proximity between point sets, offering consistency to sampling density variation. It is computed as:

$$\text{CD}(P, Q) = \frac{1}{|P|} \sum_{p \in P} \min_{q \in Q} \|p - q\| + \frac{1}{|Q|} \sum_{q \in Q} \min_{p \in P} \|q - p\| \quad (4.20)$$

Providing a symmetric average of nearest-neighbor discrepancies. CD is particularly useful for assessing local fidelity in heterogeneous or sparsely sampled regions.

The Hausdorff Distance (HD) captures the worst-case mismatch between point sets:

$$\text{HD}(P, Q) = \max \left\{ \sup_{p \in P} \inf_{q \in Q} \|p - q\|, \sup_{q \in Q} \inf_{p \in P} \|q - p\| \right\} \quad (4.21)$$

Highlighting outliers and localized reconstruction failures. HD is sensitive to isolated errors and is used to detect structural inconsistencies in occluded or poorly reconstructed regions.

Together, these metrics form a layered evaluation framework: RMSE captures global bias, CD reflects average local accuracy, and HD exposes extreme deviations. For consistency, all nearest-neighbor queries used in CD and HD computations employ the same spatial indexing and metric definitions as those used during integration and feature clustering.

Figure 15 provides a visual summary of these metrics, illustrating their conceptual differences and diagnostic roles.

## 4.5.2 Clustering validation

Clustering evaluation relies on internal measures that reflect cohesion and separability. The Silhouette index per sample is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (4.22)$$

where  $a(i)$  is the mean intra-cluster distance for sample  $i$  and  $b(i)$  the minimum mean distance to points in any other cluster. The Davies–Bouldin index aggregates cluster-level dispersion:

$$\text{DB} = \frac{1}{K} \sum_{i=1}^K \max_{j \neq i} \left( \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right), \quad (4.23)$$

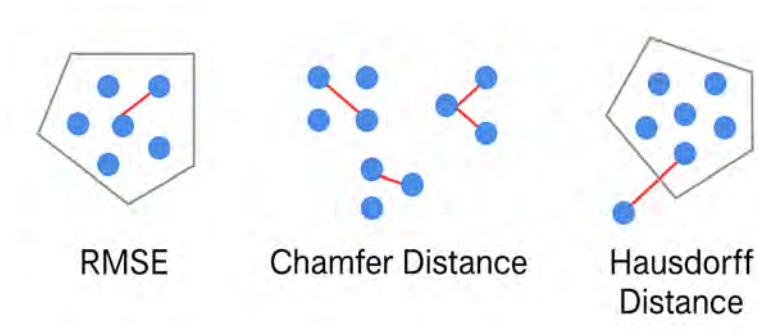


Figure 15 – Illustration of geometric evaluation metrics. Left: RMSE measures global deviation. Center: Chamfer Distance captures average proximity. Right: Hausdorff Distance highlights worst-case mismatch.

with  $\sigma_i$  within-cluster scatter and  $d(c_i, c_j)$  centroid distance. Lower DB values indicate improved separability.

### 4.5.3 Computational metrics and timing analysis

Operational viability is assessed through a timing and resource-usage protocol. End-to-end execution time is decomposed by pipeline stage and normalized per million points for scene-size extrapolation. Memory footprints and GPU utilization are recorded; feature clustering throughput (points processed per second) is reported for representative scene sizes. When algorithmic nondeterminism affects timings, statistics are expressed as means with standard deviations over repeated runs with fixed seeds.

### 4.5.4 Experimental and comparative procedures

Comparisons isolate information loss and evaluate integration strategies. Cross-evaluation tests perform bidirectional comparisons between integrated outputs and original sensor clouds to detect geometric drift, attribute degradation or completeness gain. Integration comparison experiments assess pairwise and full multimodal integration strategies. All experiments are scripted in a reproducible codebase built on Open3D and NumPy; configuration files and seeds are archived to enable exact reruns.

## 4.6 Computational Architecture and Implemented Algorithms

The project integrates coordinated algorithms for acquisition, processing, feature extraction and visualization. The solution is organized into macro-processes: multimodal integration and alignment, deep-learning feature extraction, and continuous rendering via Gaussian-based splatting. The pipeline is implemented in Python, with critical rasterization kernels accelerated in CUDA/C++.

### 4.6.1 Integration and alignment of point clouds

The pipeline begins with Integration and alignment. Where GPS coverage is uneven, a manual pairwise anchoring alignment is applied: two corresponding reference points are selected per cloud; the mean  $Z$  is subtracted to normalize the horizontal plane; and when angular deviation between reference vectors exceeds  $1^\circ$  a corrective rotation is applied. Quantitative data integration validation uses RMSE, Chamfer and Hausdorff metrics computed as described above.

### 4.6.2 Integration and alignment of point clouds

The pipeline begins with integration and alignment of modality-specific point clouds. Where GPS coverage is uneven, we apply a manual pairwise anchoring alignment: two corresponding reference points are selected per cloud; the mean  $Z$  is subtracted to normalize the horizontal plane; and when angular deviation between reference vectors exceeds  $1^\circ$  a corrective rotation is applied. A concise algorithmic description:

---

**Algorithm 1:** Reference-pair multimodal alignment

---

**Input:** RGB, THERMAL, LIDAR point clouds (.ply)

**Output:** Aligned point clouds

**For** *cloud* **Do**

    | Select 2 corresponding points manually;

    | Subtract mean  $Z$  from the cloud;

Apply translation between corresponding pairs;

If angular deviation, apply corrective rotation;

Save as \*\_final.ply;

---

Integrated clouds are concatenated and exported; we compute quantitative metrics (RMSE, Chamfer, Hausdorff) for validation:

### Feature extraction and encoder comparison

From each integrated multimodal cloud we evaluate several encoder approaches: PointNet, PointNet++, DGCNN, KD-tree based procedures and Point-MAE. Extracted

---

**Algorithm 2:** Integration and quantitative evaluation

---

**Input:** Aligned clouds  
**Output:** Integrated clouds + metric tables  
**For** *combination* **Do**  
  Concatenate points and attributes;  
  Export FUSION\_.ply;  
**For** *test batch* **Do**  
  Compute RMSE, Chamfer, Hausdorff;  
  Export tables;

---

embeddings are combined as in Equation 4.14 and clustered to infer region partitions. Standard supervised metrics are used when ground truth exists; for unsupervised experiments we rely on internal cluster metrics.

### 4.6.3 Rendering with Gaussian-based splatting

For continuous rendering, the pipeline adopts Gaussian-based splatting as described in Equation 4.10, with compositing defined by Equation 4.11. In this approach, each point in the integrated cloud is represented as an anisotropic 3D Gaussian, whose spatial support, opacity and multimodal attributes contribute to a smooth and continuous rendering.

The rasterization process employs soft z-buffering to approximate occlusion and is implemented using GPU-accelerated kernels for performance. Unlike fully differentiable Gaussian splatting pipelines, the current implementation does not perform multi-view optimization or backpropagation of rendering loss. As a result, the rendered outputs should be interpreted as viewpoint, dependent continuous rendering than fully optimized neural renderings.

The rendering loop follows a structured sequence, initializing splats from the annotated point cloud and compositing them from each camera pose. Although the pipeline supports differentiable rasterization, the Gaussian parameters  $(\mu_i, \Sigma_i, \alpha_i)$  are not refined through gradient descent in this version. The rendering loop is summarized below:

---

**Algorithm 3:** Gaussian-based splatting rendering loop

---

**Input:** Annotated point cloud  
**Output:** Continuous rendered scene  
Initialize Gaussians (centers, covariances, opacities);  
**For** *camera pose* **Do**  
  Project splats into image plane;  
  Rasterize with soft depth ordering;  
  Composite contributions per pixel;  
Export rendered sequence and latent fields;

---

This rendering strategy enables smooth visualization of multimodal attributes (e.g., RGB and thermal) while preserving geometric continuity. It also facilitates integration with downstream analysis and clustering modules, as the splatted representation maintains spatial coherence and attribute consistency across viewpoints.

For full implementation details, parameter tables and the flowchart of the Gaussian module, see Appendix C.

## Conclusion

This chapter presented the integration and implementation details of the `GaussianFusion_IA` computational architecture. Starting from multimodal data acquisition (RGB, thermal, LiDAR) we described preprocessing, registration and integration steps that ensure geometric and radiometric integrity across sensors. Gaussian-based splatting was introduced as the continuous modeling backbone (Eq. 4.10), enabling smooth surfaces enriched with sensor attributes.

We detailed feature extraction using a hybrid of PointNet++, DGCNN and PointMAE to produce robust embeddings (Eq. 4.14) subsequently clustered by k-Means or DBSCAN (Eqs. 4.16, 4.17). Continuous viewpoint-dependent rendering completes the pipeline, allowing visualization and integration with analysis stages.

The overall flow (Figure 12) ties acquisition, integration, feature clustering and rendering into a modular, scalable system that supports quantitative and qualitative evaluation. Appendix B documents the Gaussian-based splatting implementation, parameter values and comparative experiments that motivated the terminology and design decisions adopted in this chapter.

## 5 MULTISENSORY EXPERIMENTS

This chapter details the practical and technical foundations that enabled the collection of multimodal data used in experiments with the `GaussianFusion_AI` architecture. It provides the methodological groundwork that precedes and supports the results presented in the following chapters, clearly specifying which sensors were used, how they were mounted, which flight strategies were adopted, and how the data were prepared for the integration and inference stages.

Unlike exclusively synthetic approaches or experiments based on pre-existing datasets, this research is built on real data acquired in distinct urban scenarios, each with its own degree of geometric, structural and semantic complexity. Field campaigns required prior planning, sensor calibration, automated execution and continuous geospatial control with real-time correction. The success of this stage had a direct impact on the fidelity of the generated 3D models and on the quality of the inferred clusters.

The sections that follow first present the physical and computational infrastructure used, including processing stations, aerial platforms and the onboard sensors. Then each urban dataset is treated individually with attention to its topographic, sensory and operational characteristics. Presenting the material in this sequence offers an integrated view of the pipeline from capture to analysis and clarifies the method’s limitations and capabilities in the territory.

### 5.1 Computational Infrastructure and Aerial Platforms

Data acquisition and processing were supported by two high-performance workstations configured with Intel Core i7-12700KF processors (12 physical cores), 64 GB DDR5 RAM and NVIDIA RTX 4070 GPUs with 12 GB VRAM. The systems ran Windows 11 and a software stack optimized for deep learning, three-dimensional rendering and point-cloud processing; all heavy computations, from unsupervised embedding extraction to continuous rendering with Gaussian-based Splatting, were executed on these workstations, exploiting CPU/GPU parallelism and high-speed NVMe storage for intermediate datasets and caches.

For aerial capture, two DJI platforms were used: the Matrice 350 RTK and the Mavic 3T Enterprise. The Matrice 350 RTK operated as the primary technical platform for high-fidelity surveys and carried the Zenmuse L2 hybrid sensor, while the Mavic 3T Enterprise was employed for rapid or access-limited missions. The Matrice 350 RTK integrates an airframe capable of carrying heavy payloads and a built-in RTK receiver enabling centimetric positioning when paired with a local base station. The Mavic 3T Enterprise provides a compact alternative with higher optical resolution and integrated

radiometric thermal sensing, suitable for façade inspection and thermal anomaly detection in constrained environments.

Figure 16 shows both aircraft side by side at comparable scale to highlight their details. The figure was obtained directly from the manufacturer’s website (3). Mission planning and execution relied on DJI Pilot 2, which automates route generation with configurable strip overlap, speed, camera angle and spacing. Flight altitudes were chosen between 40 and 80 meters according to terrain, occlusion and desired ground sampling distance, and all captures included real-time geospatial correction via an external RTK antenna synchronized with the airborne receiver. A representative mission planning interface is shown in Figure 17, configured for full multisensor coverage over the CTEEx area.



Figure 16 – DJI Matrice 350 RTK and Mavic 3T Enterprise, used in outdoor experiments. Source: (3).

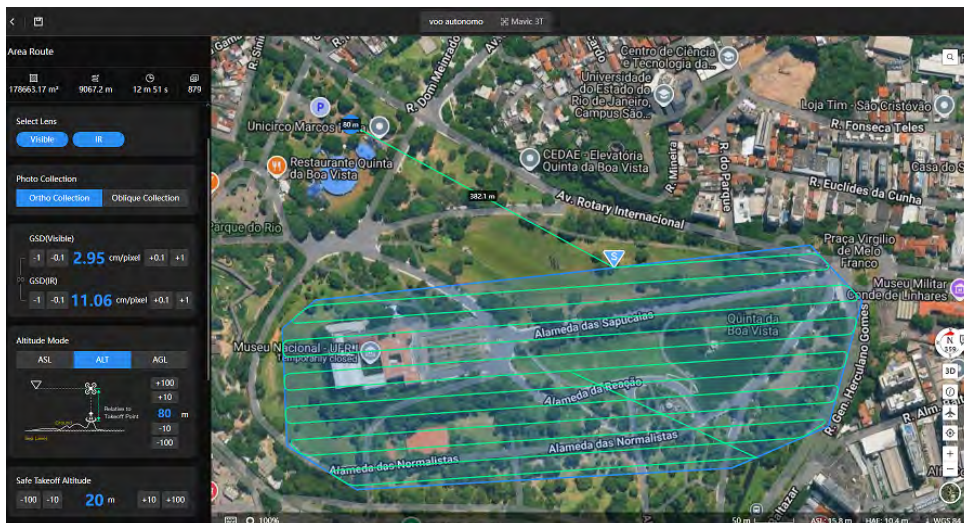


Figure 17 – DJI Pilot 2 interface used for automated multisensor mission planning and route configuration for the CTEEx surveys. Adapted from manufacturer documentation (4)

The Matrice 350 RTK platform is documented in Figures 18 and 19, which provide complementary annotated views highlighting external and internal elements relevant to field operations and data provenance. Figure 18 depicts the Matrice with labeled propellers, motors, collision sensors, landing gear, battery compartments and the RTK antenna. The configuration of motors and propellers determines available thrust and

control authority, while collision sensors and landing gear affect operational safety and payload clearance; battery placement and hot-swap accessibility define practical endurance and turnaround time between flights. Figure 19 focuses on the Unit Control module, the locus of avionics and time-synchronization: inertial sensors (IMU), barometer, magnetic compass, GNSS/RTK receiver, the autopilot and flight controller, the electronic speed controllers (ESCs), video transmitter and the accessory connector that supplies power and data to payloads. Locating these components clarifies the hardware boundaries where sensor integration, time stamping and real-time corrections are performed, and explains sources of latency and drift that propagate into SfM tie points and LiDAR georeferencing.



Figure 18 – DJI Mavic 3T Enterprise — annotated external components: propellers and motors, collision sensors, landing gear, battery compartments and RTK antenna. Source: Adapted from (3)



Figure 19 – DJI Matrice 350 RTK — annotated external components: propellers and motors, collision sensors, landing gear, battery compartments and RTK antennas. Source: Adapted from (3)

The Zenmuse L2 payload carried on the Matrice combines a stabilized 4/3" 20 MP RGB camera and a pulsed LiDAR scanner and is documented in Figure 20. The L2's gimbal

stabilization, optical aperture and LiDAR aperture are annotated to connect hardware elements with their operational roles: radiometric capture, stabilization and laser ranging. The L2's LiDAR supports up to three returns per pulse and throughputs that enable dense sampling under typical survey altitudes; when coupled with integrated RTK corrections and the aircraft's inertial measurements, the L2 yields dense, geometrically consistent point clouds with centimetric absolute accuracy suitable for urban reconstruction and classification tasks.



Figure 20 – Zenmuse L2, hybrid sensor used for integrated RGB and LiDAR acquisition.  
Source: (3)

Field georeferencing used a local RTK base station to provide differential corrections to the airborne receivers; the base's GNSS antenna, radio modem and power module are annotated in Figure 21. Establishing a stable link between the base station and airborne RTK receiver was an operational requirement for achieving the centimetric absolute accuracies reported in the reconstruction evaluation.

Table 2 summarizes the principal technical characteristics of the two platforms and their payloads, while Table 3 enumerates the sensors, the platform that carried them and the collection sites where they were deployed. The complementary combination of LiDAR-enabled RGB capture from the Matrice+L2 and high-resolution optical plus radiometric thermal imaging from the Mavic 3T Enterprise enabled evaluation of the `GaussianFusion_AI` pipeline across a spectrum of input conditions, from dense geometric coverage to scenarios where optical and thermal cues dominate.

Mission planning, flight execution and payload control procedures were standardized across datasets. Automated flight plans were executed with conservative overlap and sidelap settings to guarantee redundant viewpoints for SfM tie-point stability and dense MVS



Figure 21 – RTK base station with GNSS antenna, radio modem and power module; used to provide differential corrections to airborne RTK receivers. Source: From (3)

Table 2 – Technical comparison of the drones used

Specification	Matrice 350 RTK + L2	Mavic 3T Enterprise
RGB sensor	20 MP (4/3")	48 MP (1/2") + 12 MP tele
Thermal sensor	–	640×512 px, radiometric
LiDAR	Yes (3 returns)	No
RTK	Integrated	External (D-RTK 2)
Endurance	55 min	45 min
Total weight	6.5 kg	920 g

Table 3 – Sensors and corresponding acquisition areas

Sensor	Platform	Capture modality	Collection sites
Zenmuse L2	Matrice 350 RTK	LiDAR (3 returns) + RGB	CTEx, Quinta da Boa Vista
RGB camera	Mavic 3T	Optical imagery (48 MP)	CTEx, Glória, Quinta da Boa Vista
Thermal camera	Mavic 3T	Infrared (640×512 px)	CTEx, Glória, Quinta da Boa Vista

reconstruction; LiDAR scanning parameters were configured to balance point throughput and surface penetration for complex urban canopies. All sensor time stamps and logs were archived to permit offline verification of synchronization and to support post-hoc reprojection and residual analysis.

## 5.2 CTEEx Dataset — Integration and Feature Clustering Evaluation in a Controlled Field

The CTEEx dataset (Centro Tecnológico do Exército) in Guaratiba, RJ, served as the initial validation site for `GaussianFusion_AI`. Its topographic and spectral configuration offers a favourable testbed: structural simplicity combined with controlled metric density. The scene consists of a flat field with low vegetation, a few metal technical structures and two goalposts. Lacking elevation changes, architectural interference or complex shadows, this terrain is suitable to isolate sensor behaviour and validate multimodal data integration and latent feature clustering.

Unlike urban or forested areas where visual complexity can obscure alignment errors or geometric inconsistencies, CTEEx functions as an open three-dimensional laboratory. The absence of significant occlusion, combined with scene symmetry, permits direct observation of structural properties extracted from different sensory channels.

### 5.2.1 Acquisition and Sensor Reconstruction

Data were acquired across two independent flight trials with different platforms and complementary sensors. The first trial used the Mavic 3T Enterprise with a 48 MP RGB camera and a radiometric thermal camera ( $640 \times 512$  px). Both sensors were synchronized and captured 169 images per sensor with a minimum overlap of 85%, mean altitude of 45 m and a double-grid flight pattern to ensure homogeneous angular coverage.

The second trial used the Matrice 350 RTK equipped with the Zenmuse L2 LiDAR. The LiDAR provides multi-return capability, an internal IMU, GNSS and real-time correction via an NTRIP connection, yielding dense clouds with metric precision below 5 cm, validated in the literature (154). Average point density exceeded  $800 \text{ points}/\text{m}^2$  in central areas.

RGB and thermal clouds were reconstructed using SfM and MVS adapted for multispectral imagery. Thermal reconstruction required additional radiometric calibration and rejection of low-emissivity frames. The LiDAR cloud was processed from native LAS files with return filtering and flight surface adjustment.

Figure 22 shows sensor-specific clouds and pairwise integrations; zoomed crops over a goalpost and a technical box facilitate visual comparison. The RGB cloud provides detailed texture but exhibits depth imprecision at lateral edges. The thermal cloud offers emission patterns useful for surface analysis but is sparse over neutral surfaces. The LiDAR cloud preserves structural fidelity with stable alignment and no geometric duplication.

Multisensory compositions were evaluated using geometric distances between clouds (RMSE, Chamfer, Hausdorff) and preserved point counts.

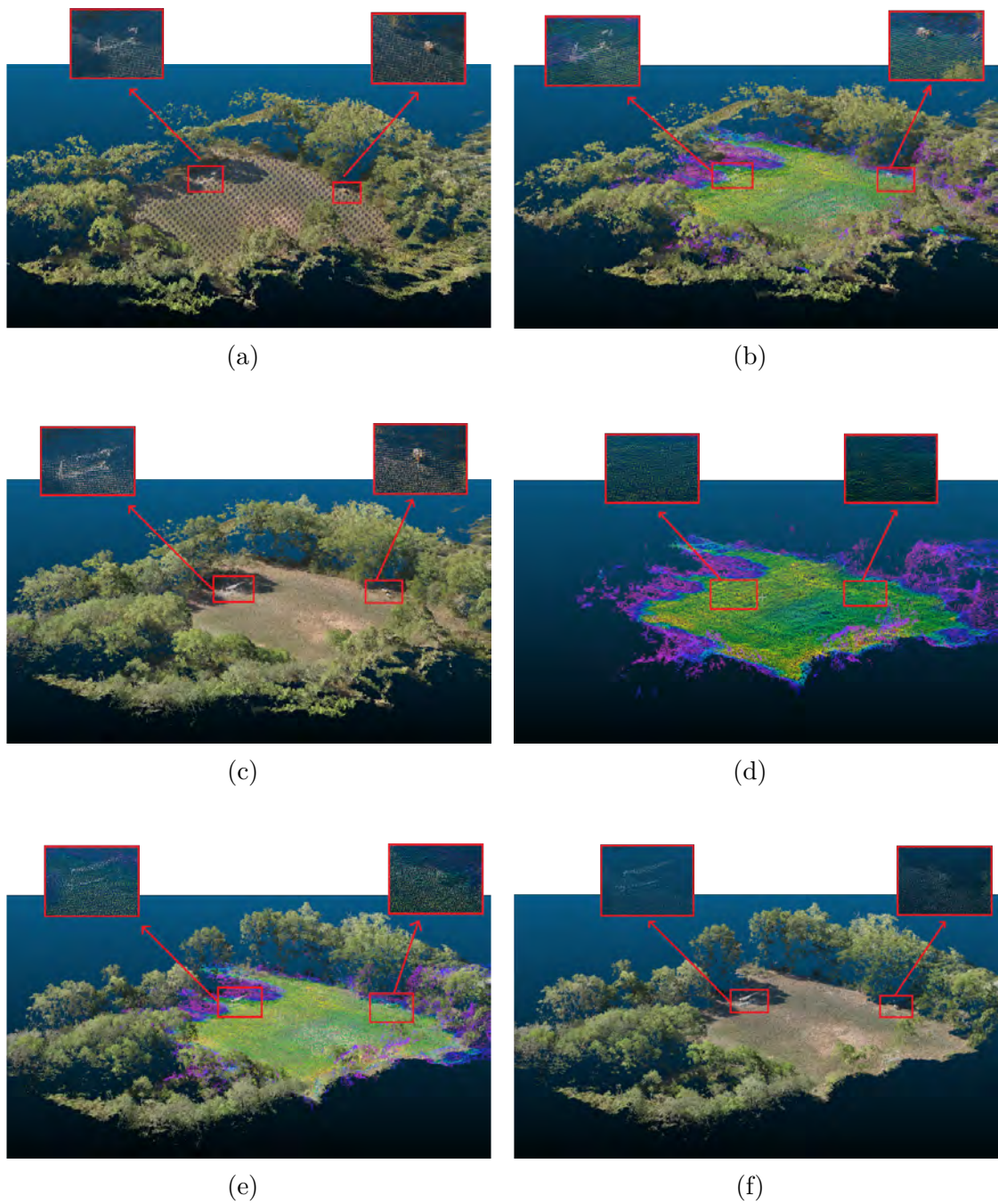


Figure 22 – Per-sensor reconstructions and pairwise integrations for the CTEEx dataset, focused on a goalpost and a heat-emitting box. (a) RGB-only cloud. (b) RGB + thermal integration. (c) RGB + LiDAR integration. (d) Thermal-only cloud. (e) Thermal + LiDAR integration. (f) LiDAR-only cloud. All views depict the same cropped region to compare the two scene elements across sensing modalities.

Table 4 consolidates the impact of each integration. The RGB–Thermal integration preserved the optical base ( $\text{RMSE} < 0.10$  m) and provided continuous radiometric attributes. The RGB–LiDAR integration achieved high structural consistency with Hausdorff below 15 m in peripheral zones. The isolated thermal channel produced the largest errors ( $\text{RMSE} > 11$  m), demonstrating limited stability as a geometric base.

Table 4 – Pairwise and integrated cloud comparisons — CTEEx dataset

Batch	Cloud 1	Cloud 2	RMSE (m)	Chamfer	Hausdorff	Points
B1	RGB	THERMAL	11.33	3.93	47.76	7.2 M
B1	RGB-Thermal Integration	RGB	0.098	0.005	4.17	7.3 M
B2	RGB	LiDAR	10.55	3.30	47.54	7.2 M
B2	RGB-LiDAR Integration	RGB	0.54	0.053	14.30	7.9 M
B3	Thermal	LiDAR	1.71	4.39	41.15	0.7 M
B3	Thermal-LiDAR Integration	LiDAR	0.69	0.075	14.94	0.84 M
B4	RGB-Thermal-LiDAR Integration	RGB	0.55	0.056	14.30	8.0 M

Table 5 compares compound integrations. The RGB–Thermal–LiDAR integration preserved structural adherence to the RGB base ( $\text{RMSE} = 0.55$  m) and showed improved Hausdorff behaviour relative to RGB–LiDAR (difference 3.77 m). Figure 23 illustrates the final triply integrated cloud, highlighting visual continuity and volumetric completeness.

Table 5 – Compound integration comparisons — CTEEx (Batch 5)

Batch	Cloud 1	Cloud 2	RMSE (m)	Chamfer	Hausdorff	Points
B5	RGB-Thermal-LiDAR	RGB-LiDAR	0.070	0.003	3.77	8.0 M
B5	RGB-Thermal-LiDAR	Thermal-LiDAR	7.43	1.73	42.67	8.0 M
B5	RGB-Thermal-LiDAR	RGB-Thermal	0.531	0.050	14.30	8.0 M

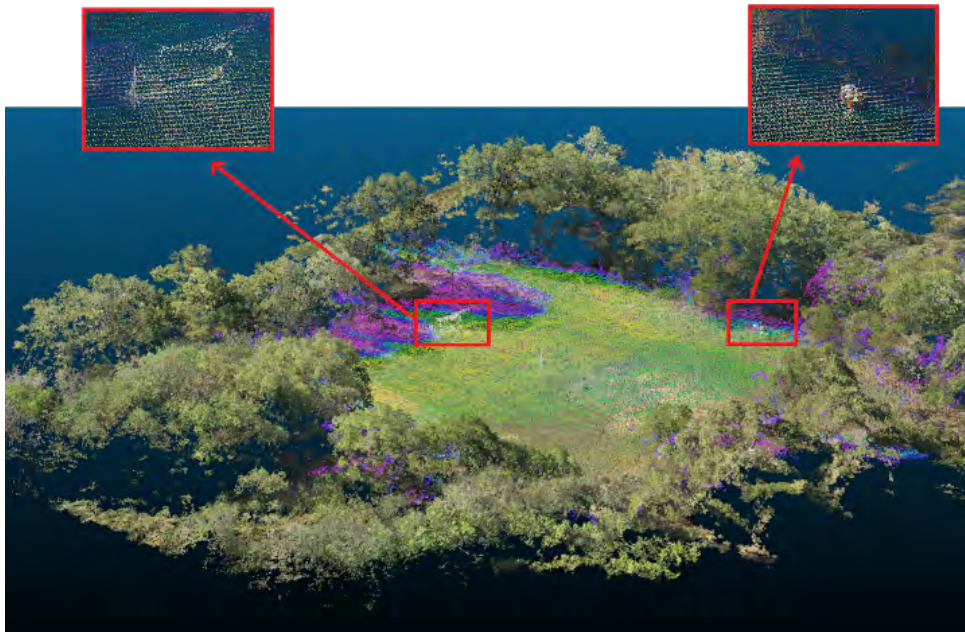


Figure 23 – RGB–Thermal–LiDAR integrated cloud for CTEEx (lateral crop).

The triply integrated cloud is a stable multisensory representation with continuous attributes and preserved geometry. Volumetry of the reference elements (goalposts, metal boxes, vegetation) was maintained, and thermal gradients enriched low-texture areas. This model is the basis for the subsequent unsupervised inference via latent feature clustering.

## 5.2.2 Unsupervised feature clustering

Unsupervised 3D feature clustering at CTE<sub>x</sub> aimed to test GaussianFusion\_AI’s ability to extract plausible structural clusters from multimodal attributes in a controlled, low-semantic-diversity environment. This step validates embedding extraction and clustering behaviour in scenes with homogeneous vegetation, flat relief and few distinct structural elements.

Although CTE<sub>x</sub> is geometrically and radiometrically stable, its low spectral and morphological variability challenges unsupervised clustering. The homogeneous grass field and few vertical structures limit latent-space separability in the absence of supervised guidance. Consequently, the clusters obtained in this experiment should be interpreted as *structural groupings*, reflecting global properties such as elevation, density and radiometric continuity, rather than semantic categories. This distinction emphasizes that the experiment was designed to verify the structural coherence of embeddings and clustering algorithms, demonstrating that GaussianFusion\_AI produces consistent structural partitions even in contexts with inherently low semantic diversity.

We tested several encoding architectures for embedding extraction: PointNet, DGCNN, Point-MAE and spatial-index approaches based on KDTree. Latent vectors were clustered using classic algorithms: k-means and DBSCAN. Table 6 summarizes performance, computation and internal quality metrics for k-means, which yielded stable and reproducible results across architectures. DBSCAN was also evaluated, but its outcomes proved highly sensitive to parameter choice ( $\epsilon$ , MinPts) and did not produce consistent clusters suitable for quantitative comparison. For this reason, only k-means metrics are reported in detail.

Table 6 – Unsupervised feature clustering performance — CTE<sub>x</sub> dataset (k-means results)

Architecture	Points	Clusters	Silhouette	DB	Time (s)	RAM (MB)	GPU (MB)	Pts/s
kdtree_kmeans	15k	6	0.3601	0.9053	3.92	212.6	0.0	3828
pointmae_kmeans	15k	6	0.1286	1.68	4.58	720.8	20.5	3272
dgcnn_kmeans	15k	6	0.0762	1.84	4.09	781.2	150.5	3663
pointnet_kmeans	150k	6	-0.0031	524.91	5.88	1044.8	151.5	25493

KDTree combined with k-means achieved the best cohesion and separability indices (Silhouette 0.36; Davies–Bouldin 0.90) and produced spatially plausible clusters: central field, peripheral vegetation and scattered technical objects. This simple configuration was efficient, memory light and CPU-only, practical for field or embedded deployments.

More complex neural encoders such as PointNet underperformed: embeddings were not discriminative, producing fragmented clusters and vertical artifacts over homogeneous surfaces. Figure 24 illustrates this behaviour—flat continuous surfaces are incorrectly split into disconnected groups, likely because the architecture struggled to capture spatial relations in a simple topology.

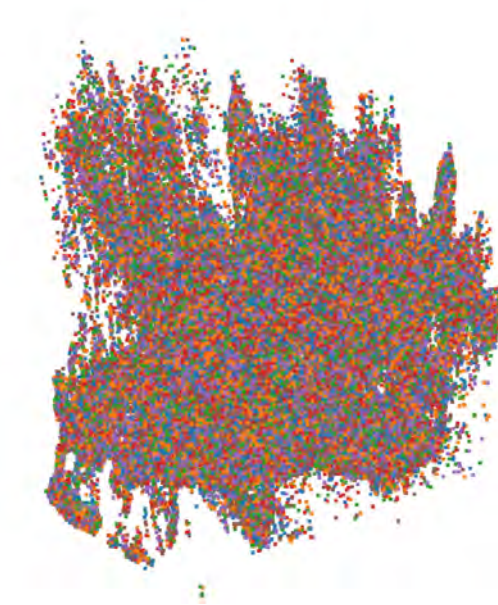


Figure 24 – Feature Clustering with PointNet + k-means. Artificial clusters and vertical distortions over flat areas.

KDTree + DBSCAN produced a conservative feature clustering: most points fell into a dominant cluster, with smaller clusters at areas of slight height variation (field edges, peripheral vegetation). This approach avoided spurious noise and preserved geometric continuity but lacked semantic expressiveness.

We also generated latent projections (UMAP and t-SNE) for KDTree embeddings. Figure 25 shows both projections coloured by attributes such as height and mean RGB intensity. Both methods reveal a diffuse central latent core with smooth branches—typical of continuous distributions with low spectral variability.

These results are expected: in visually homogeneous and structurally low-contrast environments, unsupervised feature clustering reaches natural interpretability limits. Nonetheless, the pipeline preserved important scene attributes and produced coherent projections, despite limited semantic separability inherent to the dataset.

The next chapter synthesizes CTE<sub>x</sub> findings and integrates them with broader results from `GaussianFusion_AI`.

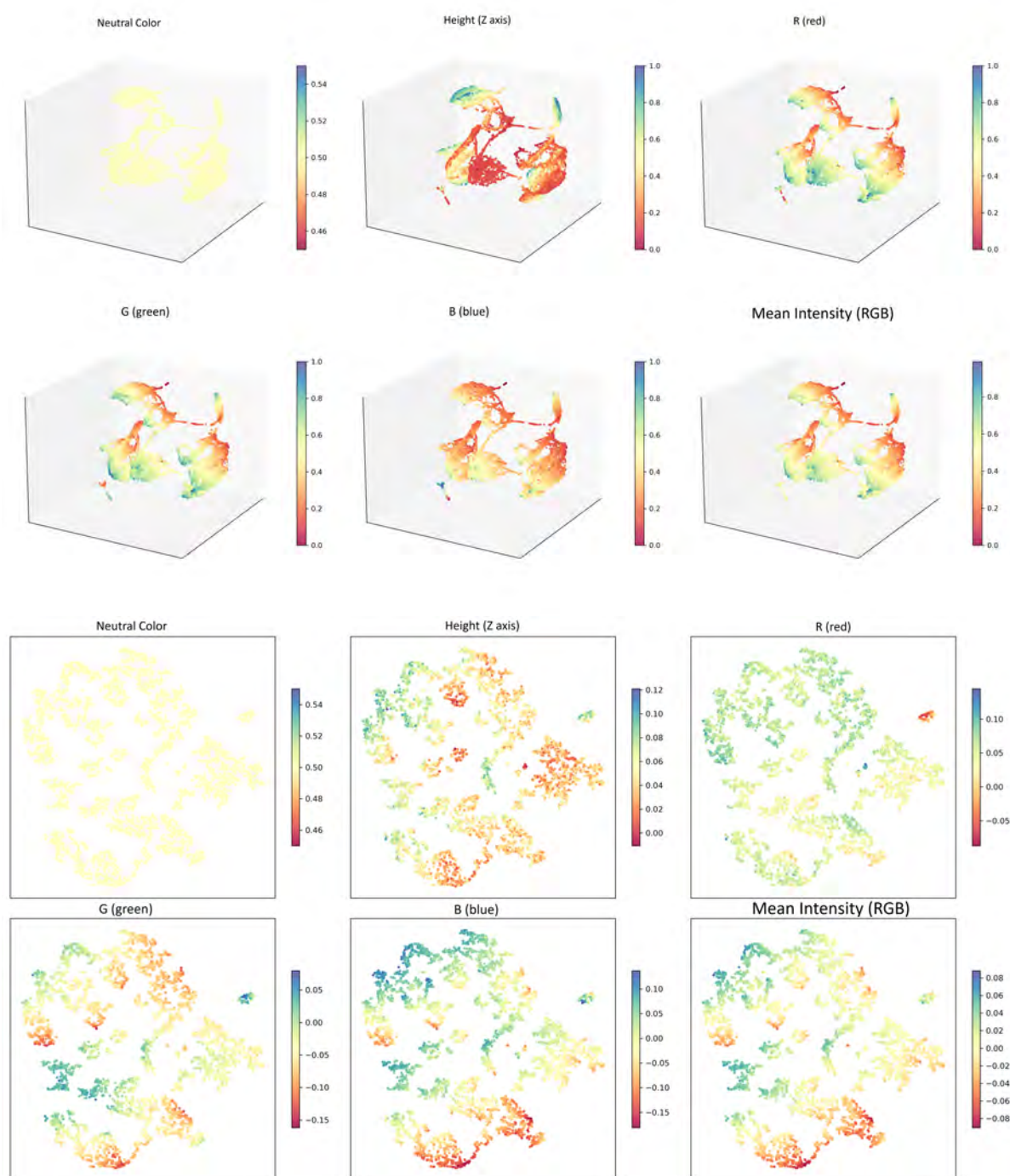


Figure 25 – Latent projections for CTEEx: UMAP (up) and t-SNE (down) coloured by attributes. Continuous core, no clearly separable clusters.

## 5.3 Gloria Dataset - Multisensory Reconstruction in an Urban Heritage Environment

The Outeiro da Glória, located above Flamengo and Glória neighbourhoods, is a culturally significant landmark in Rio de Janeiro. The Church of Nossa Senhora da Glória do Outeiro (1714–1739) and its surroundings provide a challenging testbed: irregular geometry, masonry façades, curved staircases, stone walls and peripheral vegetation. The site is listed by IPHAN since 1937 (155). Its complex relief, absence of straight lines and thermal heterogeneity make it an exemplary case to evaluate `GaussianFusion_AI` in heritage contexts where LiDAR was not deployed.

Data acquisition used only the Mavic 3T Enterprise, with 67 images per channel and perimeter flight planning on 80 meters of altitude. Optical reconstruction used SfM–MVS with thermal alignment refined by ICP. RGB–Thermal integration projected thermal data onto the dense optical mesh. This integration enabled continuous thermal gradients across façades, stairs and vegetation and supported latent semantic clustering using combined attributes.

### 5.3.1 Reconstruction and Integration Analysis

Without LiDAR, reconstruction relied entirely on imagery. Optical SfM–MVS produced a dense mesh; thermal images were aligned and interpolated onto that mesh using co-registered correspondences. The integrated RGB–Thermal cloud associates each visible point with RGB and estimated emissivity, enabling simultaneous analysis of structural and thermal variation.

Figure 26 compares: RGB reconstruction, thermal reconstruction and the integrated RGB–Thermal cloud, with a focus on the church façade.

Panels (a,b) show a well-defined RGB reconstruction with fine architectural detail. Panels (c,d) highlight limitations of thermal-only reconstruction: structural noise, low point density and fragmented silhouettes on low-emissivity surfaces. Panels (e,f) demonstrate RGB–Thermal integration: the optical geometry is preserved and enriched with continuous thermal gradients while maintaining texture and contours.

These observations underline multimodal integration’s value in urban heritage scenes. In irregular relief, light façades and contrasting vegetation, spectral and thermal combination provides complementary attributes essential for latent inference and rendering. Table 7 summarizes cross-comparisons between RGB, thermal and integrated clouds for batch GL-1.

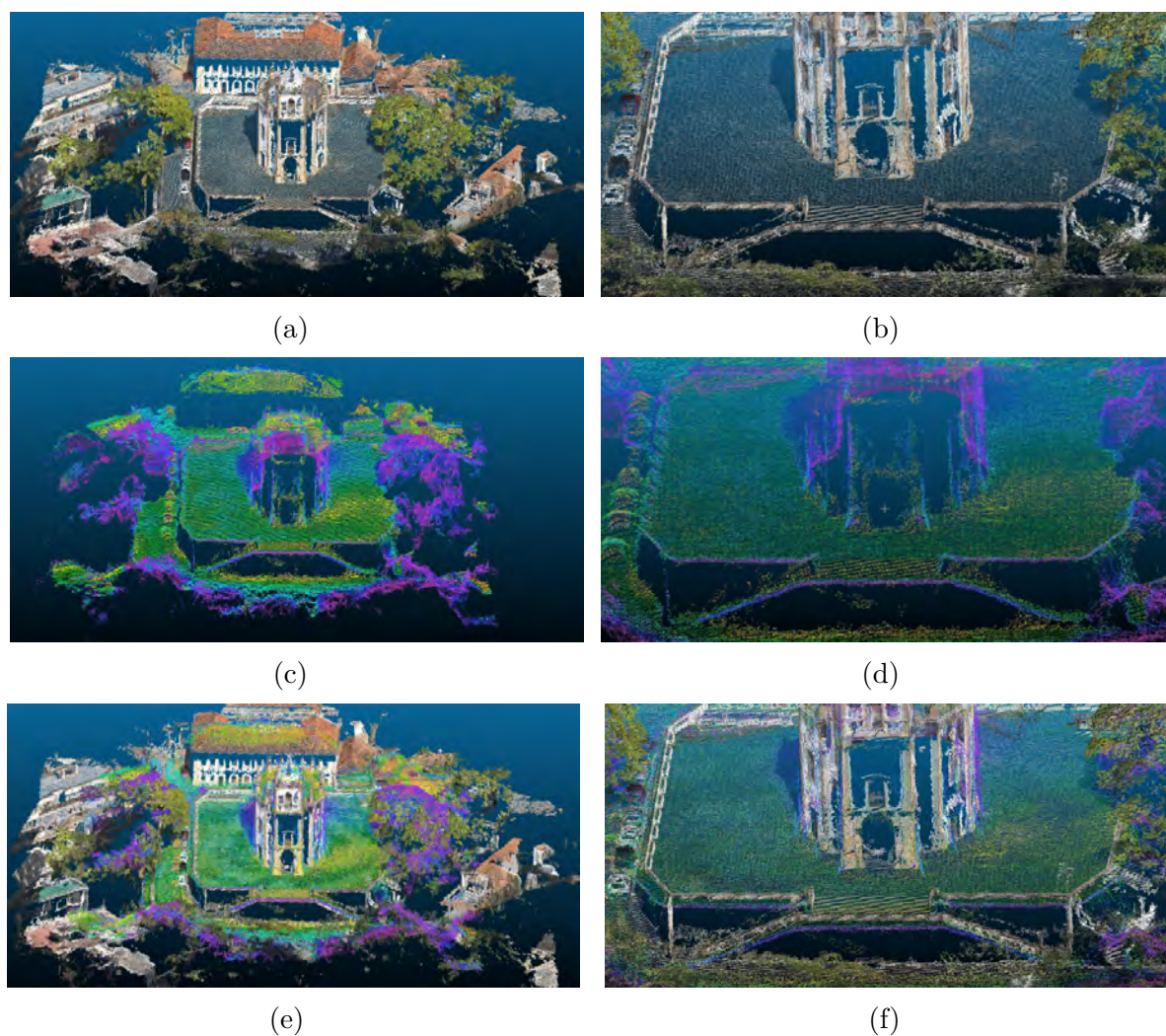


Figure 26 – Gloria reconstructions: (a,b) RGB; (c,d) Thermal; (e,f) RGB–Thermal integration (zoom on the church).

Table 7 – Comparisons for GL-1 (Gloria): RGB, Thermal and Integrations

Batch	Cloud 1	Cloud 2	RMSE (m)	Chamfer	Hausdorff
GL-1	RGB	Thermal	4.7529	1.6784	32.4485
GL-1	RGB	RGB–Thermal integration	0.0000	0.0116	7.3086
GL-1	Thermal	RGB–Thermal integration	0.0000	1.2978	32.4485

The integrated cloud preserves the RGB geometry (RMSE = 0, since integration was projected onto the optical mesh). The thermal cloud exhibits larger geometric distortion (Hausdorff > 32 m) in this urban context, underscoring that thermal data alone are often insufficient for a stable geometric reconstruction. Nevertheless, optical–thermal integration enriches semantics without compromising structural integrity when the optical base is reliable. Continuous modeling smoothed thermal gradients in shaded areas, highlighted heat retention points and detected elevation variations through emissivity changes.

Rendering the RGB–Thermal scene via Gaussian-based Splatting in a frontal pose produced a visually smooth image close to the real projection. It should be noted the result stems from interpolated texturing rather than volumetric neural reconstruction; this limitation is discussed in the next section.

### 5.3.2 Unsupervised feature clustering and Latent Projections

After constructing the RGB–Thermal integrated cloud, we performed unsupervised feature clustering to evaluate how different encoders discriminate scene regions—dense vegetation, staircases, façades and shaded zones.

We tested PointNet, Point-MAE, DGCNN, PointNet++ and hybrid encoders. Embeddings were clustered with k-means, DBSCAN and HDBSCAN. Table 8 summarizes relevant configurations: number of clusters, Silhouette and Davies–Bouldin (DB) indices, execution time, memory usage and throughput.

Table 8 – Unsupervised feature clustering performance — Gloria

Architecture	Points	Clusters	Silhouette	DB	Time (s)	RAM (MB)	GPU (MB)	Pts/s	Outliers
kdtree_kmeans	15000	6	0.3556	0.87	7.08	185.8	0.0	2117	–
pointmae_kmeans	15000	6	0.1286	1.68	7.13	695.9	20.5	2103	–
dgcnn_kmeans	15000	6	0.0762	1.84	7.71	756.8	150.5	1944	–
pointnet_kmeans	150000	6	-0.0123	117.0	8.70	1021.3	151.5	17239	–
hybrid_hdbscan	15000	2	-0.0970	13.72	6.54	696.8	16.8	2293	9759

KDTree + k-means achieved the highest Silhouette (0.36) and lowest DB (0.87) with minimal computational resources. The resulting feature clustering separated continuous surfaces (sidewalks, walls) and vegetation. Point-MAE produced richer embeddings but lower clustering performance with k-means. PointNet performed poorly, producing excessive clusters in low-contrast zones, similar to CTE<sub>x</sub>.

Figure 27 shows a 3D UMAP projection of embeddings. Cohesive regions correspond to spatially consistent clusters; branching structures reflect smooth semantic transitions such as thermal shading or vegetation density gradients. Figure 28 presents a t-SNE projection highlighting local connectivity and soft boundaries between clusters—expected in architectural textures with thermal gradients.

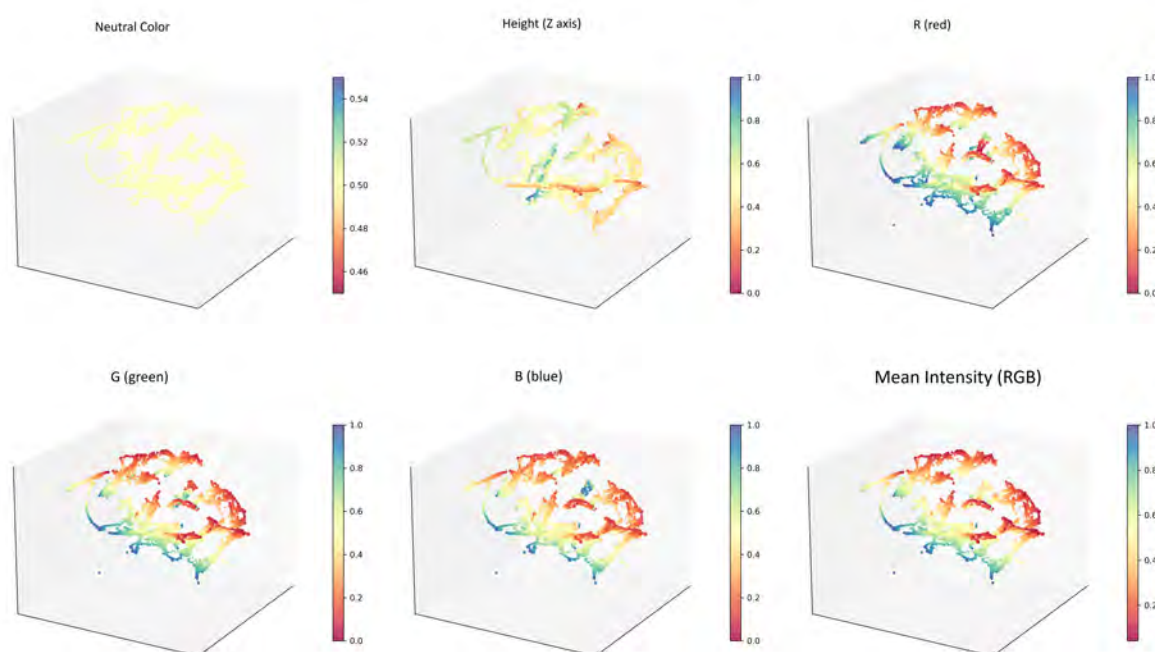


Figure 27 – 3D UMAP projection of embeddings — Gloria dataset.

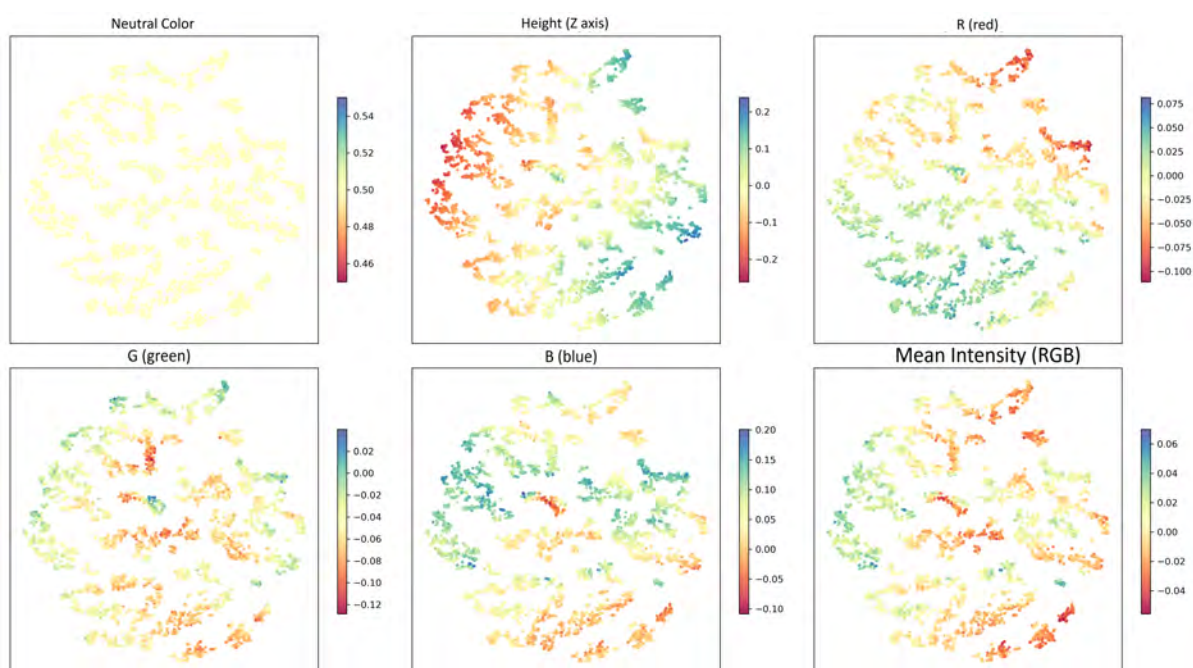


Figure 28 – t-SNE projection of embeddings — Gloria dataset.

Overall, inferred clusters respected expected structural transitions: separating vegetation from façades, isolating staircases with characteristic thermal patterns and maintaining pavement continuity. Despite the lack of a structural channel, RGB–Thermal integration provided a sufficiently stable base for latent inference.

### 5.3.3 Continuous Rendering with Gaussian-based Splatting

The technique of *Gaussian-based Splatting* was applied experimentally to the Gloria dataset to evaluate its ability to generate photorealistic images from a reconstructed multisensor point cloud in a real urban heritage environment. Recently proposed by Kerbl et al. (2), this approach is based on differentiable rasterization of three-dimensional Gaussian distributions, enabling continuous scene synthesis without requiring dense volumetric reconstruction or explicit mesh modeling.

For the initial test, a non-frontal pose was adopted, laterally centered on the church courtyard, with a three-dimensional offset defined as  $(0.0, -1.4, -0.4)$ . This configuration preserved the target object at the center of the image and kept the projection bounds  $u, v$  within the rendering plane. The resulting image, shown in Figure 29, exhibits strong visual continuity across surfaces, silhouette preservation, and absence of structural artifacts even in the presence of vertical elements such as railings and vegetation. These results demonstrate the contribution of Gaussian-based splatting: when calibrated in terms of splat density and directional focus, the technique preserves perceptual volumetry and radiometric texture of the scene, achieving coherent visualization even without solid reconstruction or explicit mesh modeling.

Subsequently, a more aggressive lateral rendering was tested using a significantly larger offset (exceeding 9000 units in the internal reference frame), surpassing the valid projection regions in image space. Although this configuration captured visually interesting peripheral zones of the scene, it compromised rendering integrity. As shown in Figure 30, partial geometry disappearance, incomplete contours, and fill failures were observed—especially in areas with low splat density oriented toward the new viewing direction. The silhouette broke at several points, assuming a fragmented appearance typical of non-convergent SfM reconstructions.

The behavior observed in these two tests highlights the operational limits of the technique: when operating within the angular space previously captured and with adequate directional distribution of splats, the method can generate continuous and photorealistic images. However, extreme pose extrapolation compromises the representation, revealing that the success of Gaussian-based Splatting depends heavily on the spatial and angular distribution of Gaussian points in the original model. This result reinforces the importance of flight plans and sampling strategies that prioritize multiple angles and coverage redundancy when robust visual rendering in diverse directions is the goal.

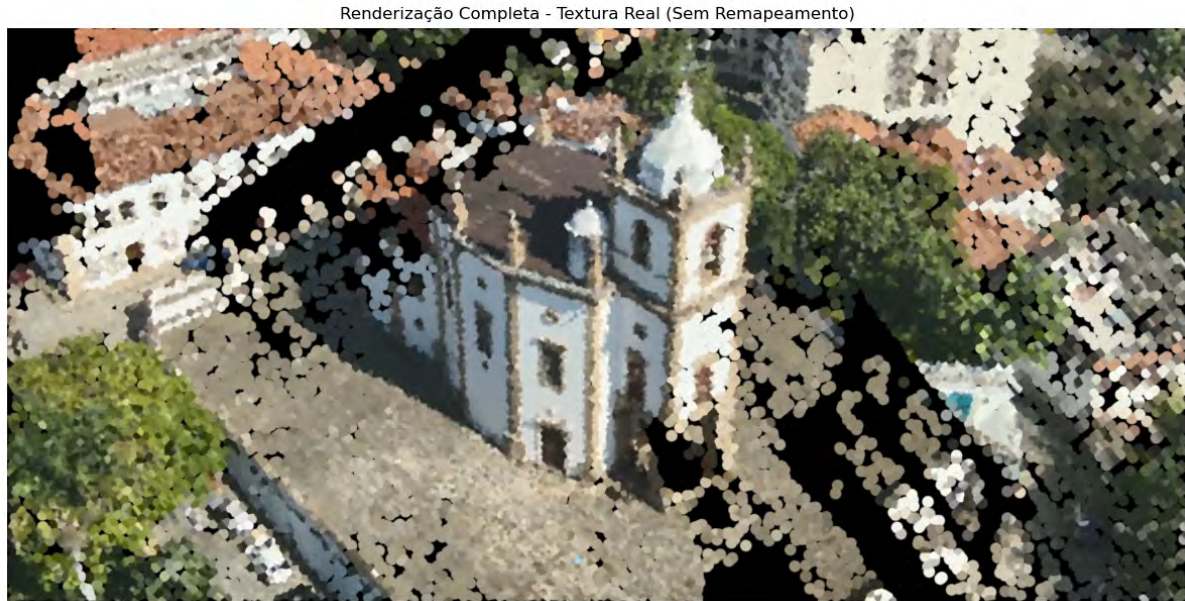


Figure 29 – Continuous rendering of the church side wall using Gaussian-based splatting. The offset expands coverage but also reveals sensitivity to viewpoint distribution, highlighting the contribution of the method in achieving photorealistic visualization while exposing its dependence on angular density.

Table 9 – Parameters used in lateral rendering

Parameter	Value	Description
MAX_POINTS	240000	Maximum number of Gaussian splats used
SIGMA	1.0	Influence radius of splats
FX_OVERRIDE	1550.0	Focal length of simulated camera
OFFSET	[0.6, -3.4, -0.4]	Three-dimensional displacement applied to the cloud

This behavior does not represent a failure of the technique, but rather a fundamental property of Gaussian-based Splatting: each projected point references the original image from its acquisition orientation. When the pose is altered and there is insufficient local density of points oriented toward the new direction, splats lack reliably assigned texture, resulting in "empty" or fragmented regions in the projected space.

Figure 31 compiles different stages of the visual pipeline applied to the Gloria dataset, beginning with an SfM-projected image, followed by a simulated 2D segmentation, a photorealistic rendering via Gaussian-based splatting, and finally a 3D feature clustering applied to the integrated RGB-Thermal cloud. This sequence demonstrates the transition from discrete image-based segmentation to continuous multimodal 3D inference, highlighting the contribution of GaussianFusion\_AI in combining clustering and rendering



Figure 30 – Continuous rendering with elevated lateral offset, showing structural coherence loss in the Gaussian representation. This illustrates the contribution of the experiment in identifying the operational limits of Gaussian-based splatting under extreme pose extrapolation.

to achieve structural coherence and enriched visualization.

This sequence reinforces three central observations:

- Direct projection of segmentations onto 2D images is fast but produces groupings inconsistent with real geometry, especially in curved or shaded zones;

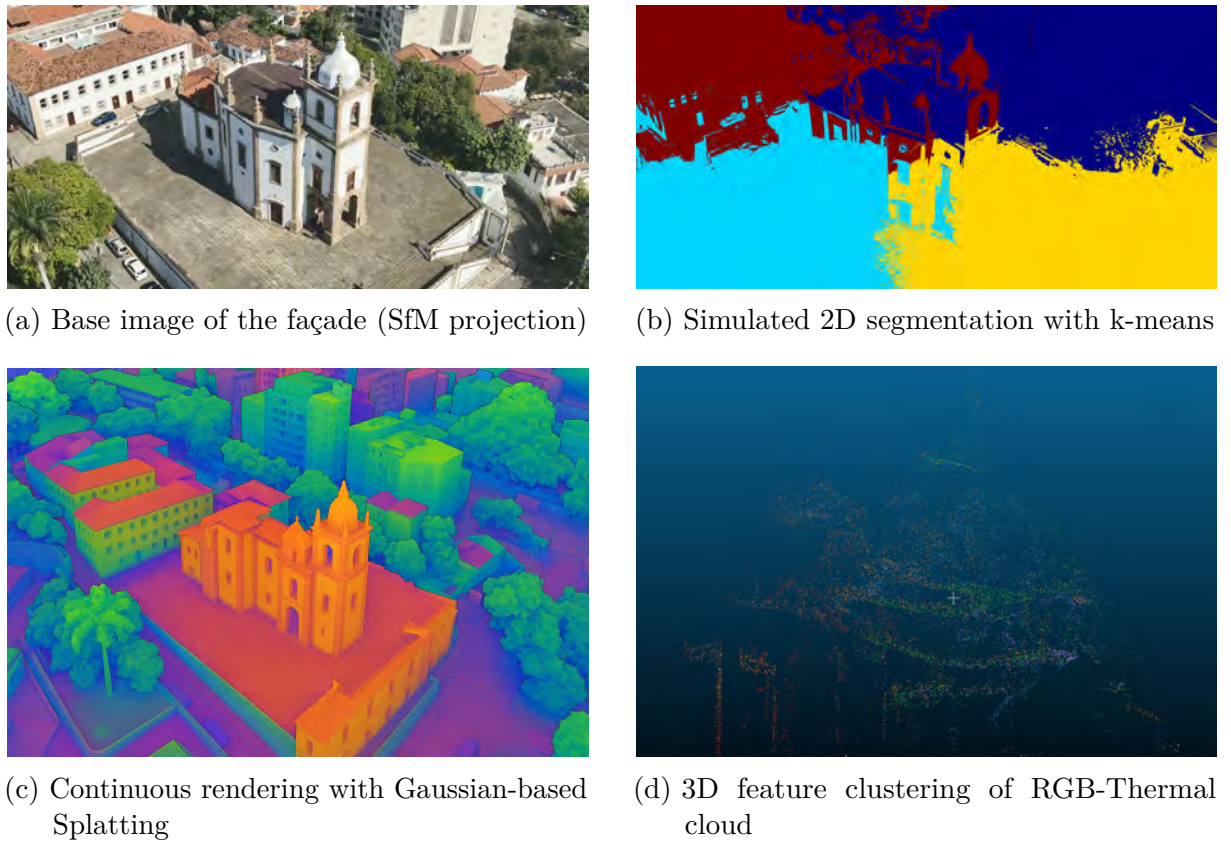


Figure 31 – Evolution of visual representation in the Gloria dataset, from SfM façade projection to simulated 2D segmentation, continuous rendering with Gaussian-based splatting, and 3D feature clustering of the RGB-Thermal cloud. The sequence demonstrates the contribution of GaussianFusion\_AI in transitioning from discrete image segmentation to continuous multimodal inference, achieving structural coherence and enriched visualization.

- Rendering with Gaussian-based Splatting approximates the scene to a coherent photographic representation, though dependent on angular density and pose calibration to avoid visual discontinuities;
- Feature clustering based on multisensor latent attributes, applied directly to the 3D cloud, enables inference of topologically plausible groupings even without mesh or supervised modeling.

Therefore, although the continuous rendering applied here remains limited to 2D projection with simulated texture, its visual results demonstrate the potential of the approach for interpretive visualization of real urban scenes. The visual fidelity obtained in the frontal pose suggests that, with greater angular density or view coverage resampling, continuous synthesis could be extended to other viewpoints—paving the way for more sophisticated experiments in the following chapter.

With the conclusion of this dataset, the research now advances to Quinta da Boa Vista, the largest and most sensor-diverse site in the thesis. There, the simultaneous application of RGB, thermal, and LiDAR channels will allow for a deeper test of the limits of the `GaussianFusion_AI` architecture in a real urban context with dense vegetation, horizontal heritage buildings, and intense structural contrasts.

## 5.4 Quinta Dataset — Integration and Feature Clustering in an Urban Park

The Quinta da Boa Vista dataset is the most heterogeneous experiment in this chapter. Unlike the controlled and relatively flat field of CTE<sub>x</sub>, Quinta represents a complex urban park with strong altitude variation, historic structures, dense vegetation, open spaces and sharply cast shadows. This diversity makes Quinta an ideal environment to evaluate the consistency of multisensor data integration and to probe the limitations of unsupervised feature clustering on richly attributed point clouds. By including this dataset, the experiments extend beyond planar contexts and explicitly test the architecture in irregular, high-complexity urban scenarios.

### 5.4.1 Acquisition and Sensor Reconstructions

Acquisitions followed the same aerial vectors as previous experiments. The Mavic 3T captured 228 RGB images and 228 radiometric thermal images (85% overlap, mean altitude 40 m). Thermal imagery was calibrated and projected onto the SfM/MVS geometry. The Matrice 350 RTK acquired LiDAR with multi-return capability, RTK correction and synchronized GNSS base.

Figures 32 and 33 illustrate the individual sensor clouds and several integration variants focusing on the National Museum façade and roof. The RGB cloud alone offers the highest point density and structural detail. The thermal cloud shows gaps, particularly over roofs and vertical elements, due to sensor resolution and emissivity limits. LiDAR presents occlusions and continuity loss in some regions caused by incidence and range constraints. Contribution: these comparisons highlight how each modality preserves specific aspects of the scene—RGB ensures geometric richness, thermal adds radiometric context, and LiDAR reinforces structural fidelity—while integration demonstrates the added value of multimodal consistency in complex urban environments.

Multimodal integrations (RGB–Thermal, RGB–LiDAR, Thermal–LiDAR and RGB–Thermal–LiDAR) yield progressively more complete clouds. RGB based integration deliver the most complete models; complete triple integration produced a spatially coherent, information-rich representation combining RGB detail, thermal radiometry and LiDAR depth.

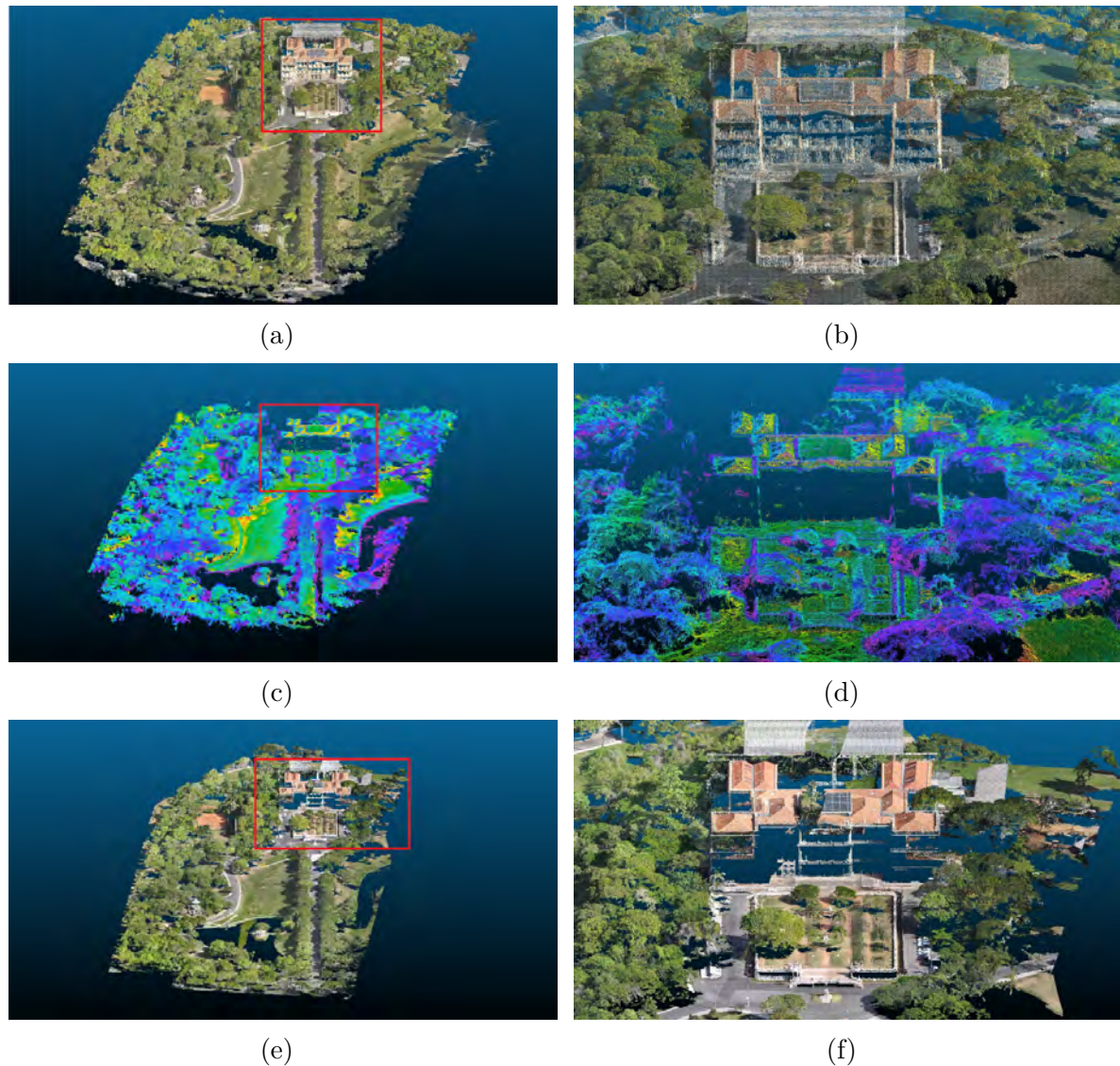


Figure 32 – Per-sensor reconstructions for Quinta da Boa Vista, focusing on the National Museum façade and roof. The comparison contributes by showing how RGB ensures geometric richness, thermal adds radiometric context, and LiDAR reinforces structural fidelity, while integration enhances multimodal consistency.

Optical-thermal reconstruction produced a high-density photogrammetric mesh with continuous thermal projection. LiDAR clouds preserved canopy, building and ground geometry. Clouds were aligned using control points and homologous features. Figure 34 shows exploratory 3D UMAP projections of latent embeddings extracted from the RGB–Thermal–LiDAR integrated cloud.

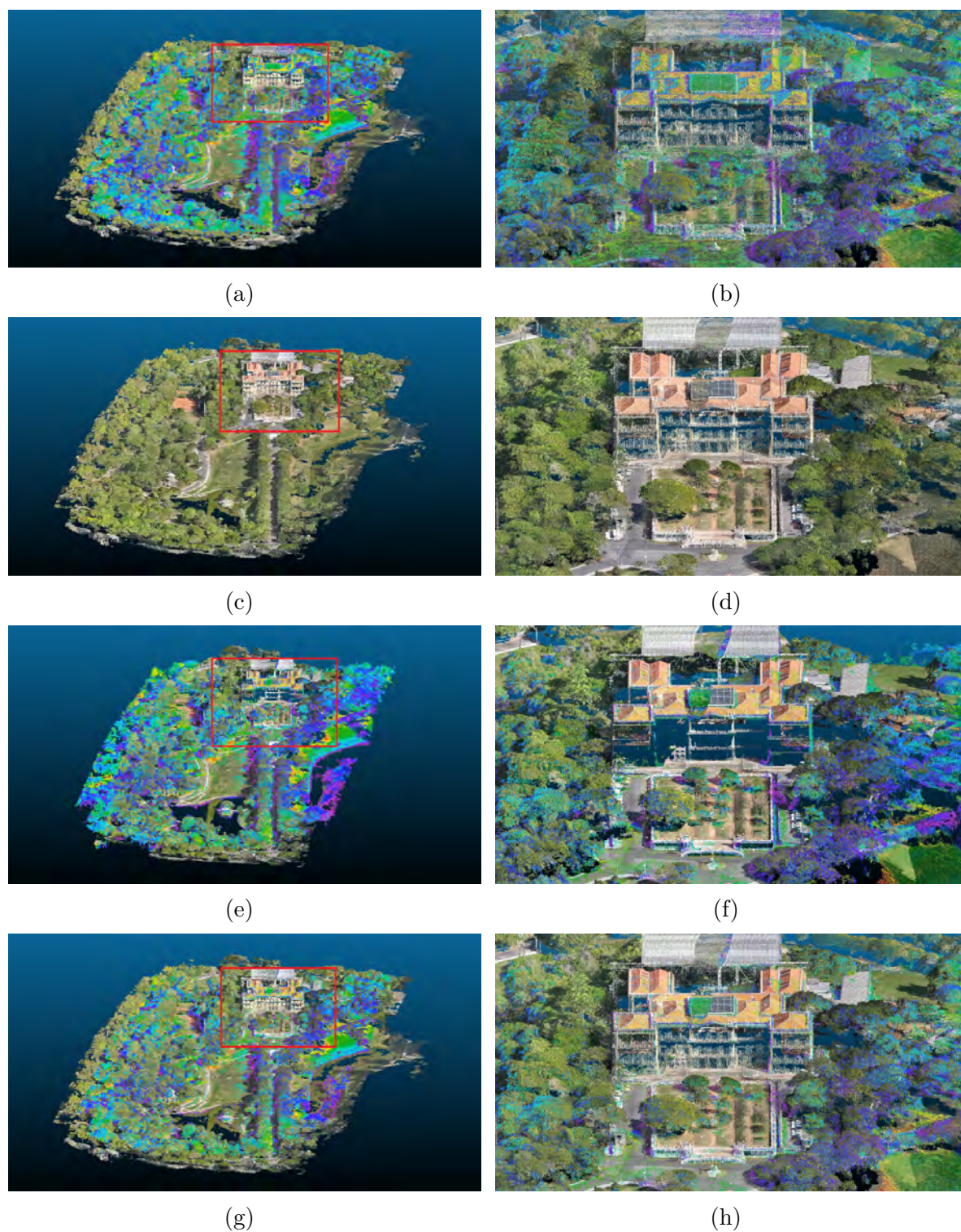


Figure 33 – Integration variants for Quinta da Boa Vista, including RGB-Thermal, RGB-LiDAR, Thermal-LiDAR, and RGB-Thermal-LiDAR (zoomed). The contribution lies in illustrating how different sensor combinations affect structural fidelity and radiometric enrichment, confirming that RGB-based integrations yield the most complete models, while triple integration achieves balanced geometric and radiometric consistency.

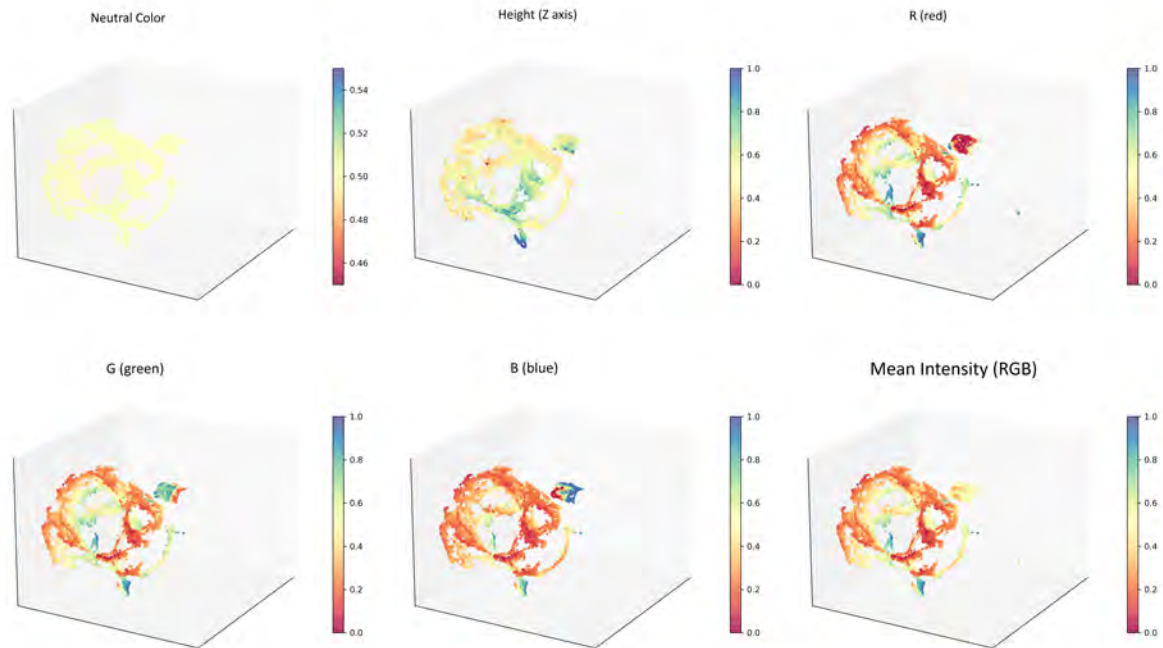


Figure 34 – Exploratory 3D UMAP projections for QUINTA, highlighting altitude, spectral channels and mean intensity.

### 5.4.2 Multisensor Integration and Quantitative Analysis

This experiment evaluated integration strategies across planar and complex vertical elements. Table 10 reports results from four initial comparison batches using RMSE, Chamfer and Hausdorff metrics.

Batch	Cloud 1	Cloud 2	RMSE	Chamfer	Hausdorff	Points
B1-Integration	RGB	THERMAL	6.7277	1.9321	80.0594	43.2 M
B1-Integration	RGB-Thermal integration	RGB	<b>0.4024</b>	<b>0.0510</b>	<b>47.6572</b>	50.3 M
B2-Integration	RGB	LIDAR	14.5950	3.6514	121.9878	43.2 M
B2-Integration	RGB-LiDAR integration	RGB	0.7716	0.1233	29.2178	63.9 M
B3-Integration	THERMAL	LIDAR	7.9578	2.8721	63.4822	20.8 M
B3-Integration	Thermal-LiDAR integration	LiDAR	<b>4.0340</b>	<b>0.4182</b>	<b>52.8473</b>	27.9 M
B4-Integration	RGB-Thermal-LiDAR integration	RGB	0.8061	0.1470	47.6572	71.2 M

Table 10 – Structural comparisons between isolated and integrated clouds — QUINTA

RGB–Thermal integration significantly reduced RMSE relative to RGB alone. RGB–LiDAR integration achieved notable geometric performance with Hausdorff 29.2 m, the lowest absolute error among optical–LiDAR comparisons. Thermal–LiDAR integration behaved intermediately: despite thermal low resolution, depth combination can be viable. The triple integration (RGB–Thermal–LiDAR) remained consistent with the RGB channel (RMSE = 0.81 m), with no major structural distortions.

It is important to emphasize that the groupings and comparisons obtained in this experiment are *structural in nature*, reflecting geometric fidelity, radiometric continuity and topographic consistency across modalities. They do not correspond to semantic categories

such as “trees” or “buildings,” but rather to coherent structural partitions that validate the consistency of multimodal integration. This distinction highlights the scientific contribution of the experiment: demonstrating that `GaussianFusion_AI` produces consistent structural groupings in complex urban environments, beyond a mere juxtaposition of techniques.

Table 11 (Batch 5) compares compound integrations directly.

Batch	Cloud 1	Cloud 2	RMSE	Chamfer	Hausdorff	Points
B5-Integration	RGB-Thermal-LiDAR	RGB-Thermal	0.6855	0.0906	29.2178	71.2 M
B5-Integration	RGB-Thermal-LiDAR	RGB-LiDAR	<b>0.2851</b>	<b>0.0270</b>	<b>47.6572</b>	71.2 M
B5-Integration	RGB-Thermal-LiDAR	Thermal-LiDAR	4.2343	0.6187	80.0594	71.2 M

Table 11 – Compound integration comparisons — QUINTA (Batch 5)

Batch 5 results indicate integration sharing the RGB sensor are the most consistent, especially when combined with LiDAR. Adding thermal data to RGB+LiDAR did not degrade geometry significantly; however, integrations lacking optical richness (e.g., Thermal–LiDAR) showed higher RMSE and Hausdorff disparity. These results reinforce that optical plus structural sensors yield the most stable 3D models and that triple sensor combinations provide robust spatial fidelity across different configurations.

## Unsupervised feature clustering

From the RGB–Thermal–LiDAR integration cloud for Quinta, we ran unsupervised feature clustering experiments with various embedding encoders and clustering algorithms. Tested encoders included PointNet, DGCNN, Point-MAE and KDTree-based structural descriptors. Embeddings were clustered with k-means, DBSCAN and HDBSCAN; samples used 15k points per run.

Table 12 summarizes results: number of clusters, Silhouette, Davies–Bouldin, execution time and resource usage.

Table 12 – Feature clustering metrics by architecture — Quinta da Boa Vista

Architecture	n_points	Clusters	Silhouette	DB	Time (s)	RAM (MB)	GPU (MB)	Pts/s	Outliers
dgcnm_dbscan	15000	4	-0.0073	169.41	11.26	794.41	150.48	1332	–
dgcnm_kmeans	15000	6	0.0156	6.79	14.72	779.47	150.48	1019	–
hybrid_dbscan	15000	4	-0.0080	99.44	13.92	777.70	16.84	1077	–
hybrid_dbscan_pca	15000	4	-0.0077	190.16	14.15	775.57	16.84	1060	–
hybrid_hdbscan	15000	2	-0.1037	13.27	14.00	719.06	16.84	1071	852
hybrid_kmeans	15000	6	0.0243	4.07	13.57	752.64	16.84	1105	–
kdtree_dbscan	15000	4	-0.0105	123.41	11.61	215.58	0.00	1292	–
kdtree_kmeans	15000	6	<b>0.3757</b>	<b>0.8582</b>	12.70	212.42	0.00	1180	–
pointmae_dbscan	15000	4	-0.0050	173.91	14.25	732.68	20.49	1052	–
pointmae_kmeans	15000	6	-0.0198	11.58	12.17	721.15	20.49	1232	–
pointnetpp	1024	1	-0.0234	43.74	11.59	902.25	40.26	88	–
pointnet_dbscan	15000	4	-0.0069	114.30	13.41	839.20	15.09	1118	–
pointnet_kmeans	150000	6	-0.0032	371.88	14.14	1044.65	151.54	10610	–

KDTree + k-means achieved the best numerical performance (Silhouette 0.3757, DB 0.8582), but visual inspection showed the high internal scores did not translate to

semantically meaningful partitions. Most of the cloud was assigned to a dominant cluster, with smaller peripheral regions associated with altitude differences, consistent with the algorithm responding to global structural properties rather than semantically defined objects such as trees, pathways or benches.

Figure 35 illustrates the results obtained with KDTree combined with k-means, where embeddings were projected with 3D UMAP, clusters were mapped back to the original cloud, and the RGB–Thermal–LiDAR input cloud was used for encoding. The unsupervised clustering yielded continuous structural partitions related to elevation, density, and emissivity, reinforcing their interpretation as structural rather than semantic groupings.

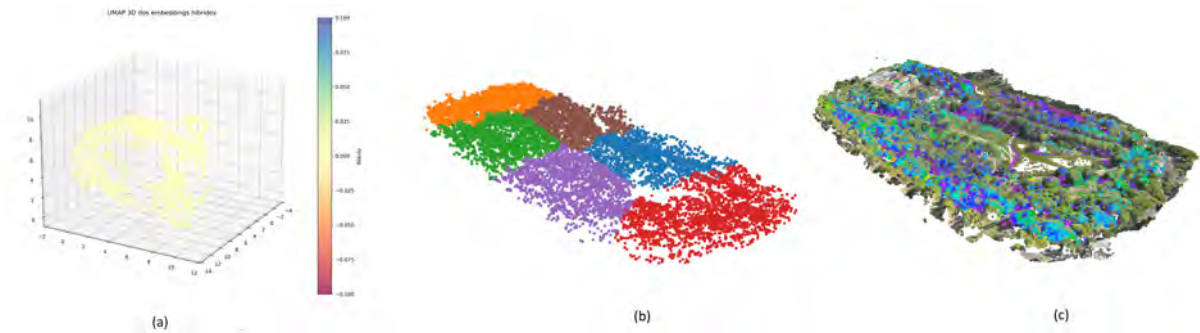


Figure 35 – Results of KDTree combined with k-means on the Quinta dataset, showing UMAP 3D embeddings, clustering mapped to the original cloud, and the RGB–Thermal–LiDAR input cloud. The contribution of this experiment is to demonstrate that unsupervised clustering yielded continuous structural partitions related to elevation, density, and emissivity, reinforcing their interpretation as structural rather than semantic groupings.

Although geometric fidelity and noise preservation are satisfactory, cluster distribution is regular and continuous, resembling spatial influence polygons; this behavior indicates that the unsupervised feature clustering responded more to global structural properties than to distinct semantic categories.

To further probe the latent structure, we produced dimensionality reductions: UMAP 2D, UMAP 3D and t-SNE. UMAP 2D (Figure 36) exhibits a dense central core surrounded by fragmented peripheral islands that may correspond to elevated areas or anomalous thermal patches; however, caution is warranted since projection artifacts can exaggerate apparent separations.

UMAP 3D preserves topology better: islands connect to the core via transitional regions, indicating continuous variation rather than clear separations (Figure 37).

t-SNE (Figure 38) further confirmed a compressed latent core with thin, poorly defined branches. Overall, latent embeddings lacked intrinsic separability sufficient to delineate real objects such as roads, vegetation or structures. This suggests the need for richer contextual inference mechanisms: deeper networks, advanced self-supervised

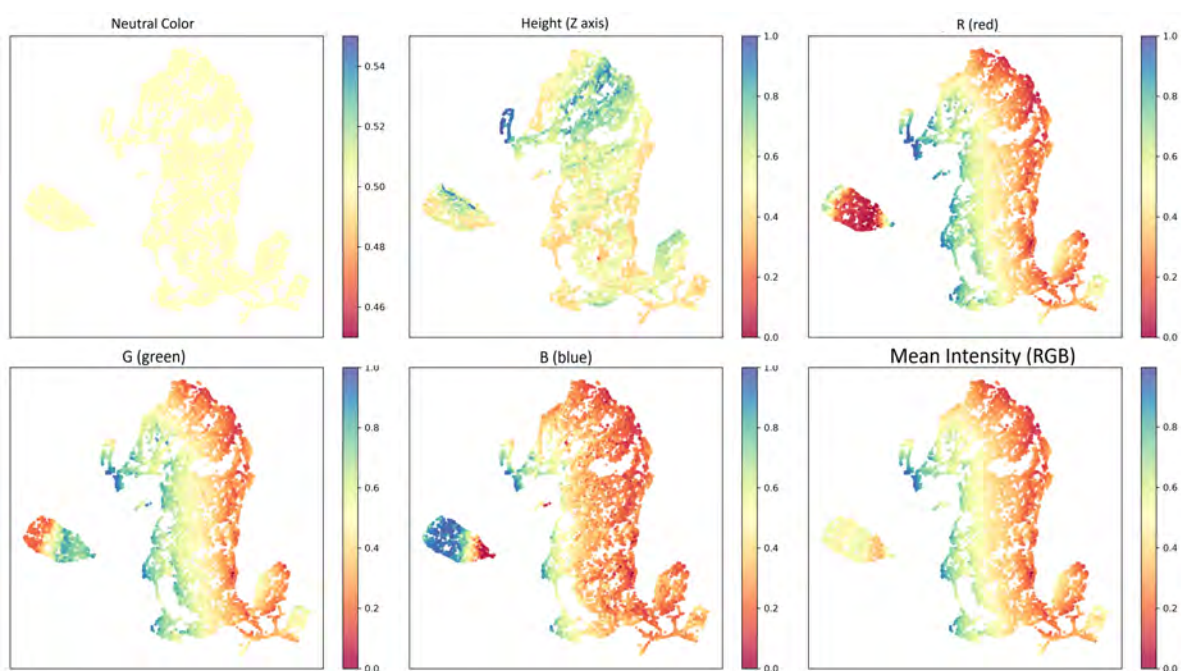


Figure 36 – UMAP 2D projection, dense core with peripheral latent islands.

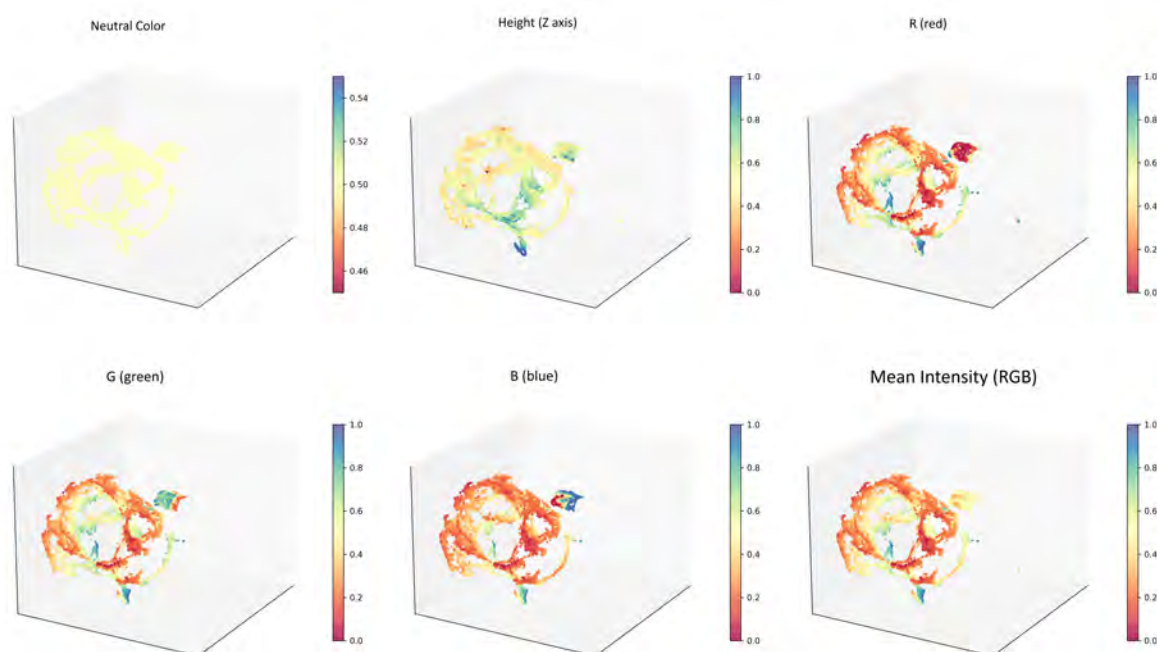


Figure 37 – UMAP 3D projection, structural continuity and absence of sharp latent cluster separations.

strategies or light supervised guidance to obtain semantically meaningful partitions in complex urban environments.

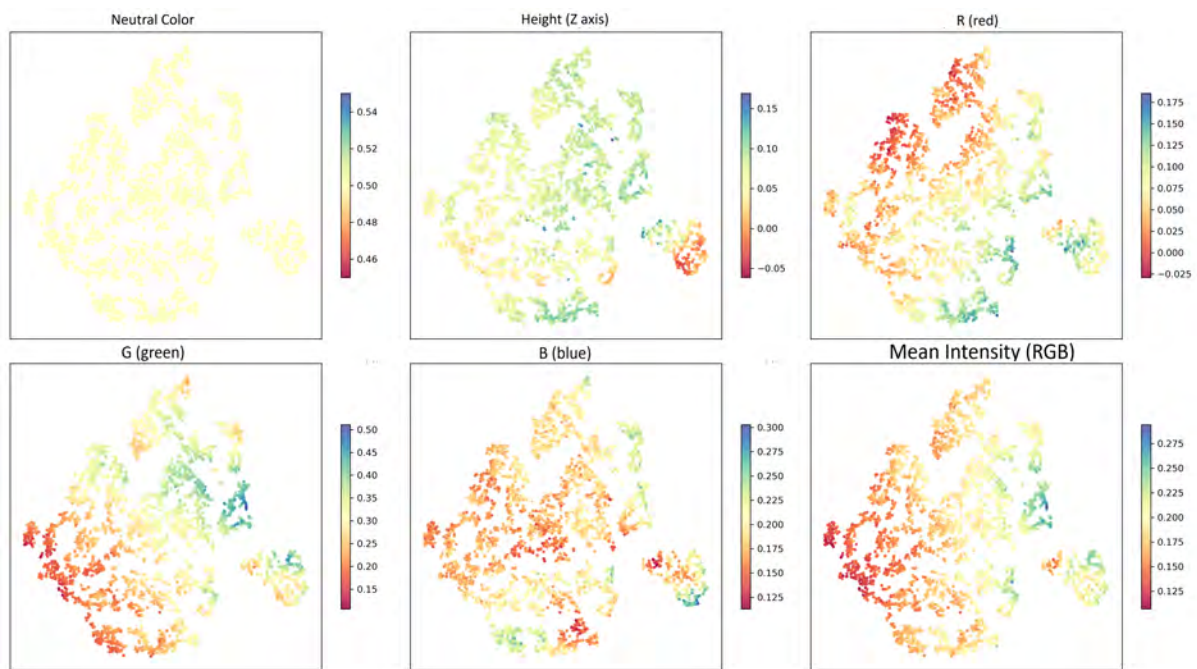


Figure 38 – t-SNE projection, compressed latent structure with thin, ill-defined branches.

## Conclusion

Experiments in this chapter evaluated `GaussianFusion_AI` across three distinct scenarios with different sensory and structural characteristics. CTE<sub>x</sub> functioned as a controlled validation field, demonstrating the consistency of multimodal integration in low-semantic-diversity contexts. Gloria introduced architectural and historical complexity and showed the value of thermal enrichment over a reliable optical base. Quinta da Boa Vista was the most challenging dataset, combining dense vegetation, large horizontal structures and pronounced morphologic variation—an appropriate test for unsupervised inference limits in real contexts.

Results confirm the pipeline’s stability in integration and reconstruction; however, semantic separation based solely on latent patterns is limited in complex urban scenes. UMAP and t-SNE projections and unsupervised clusters showed geometric coherence but also highlighted the need for methodological refinements to achieve higher-level semantic inference. These findings are discussed further in the next chapter, which offers a critical assessment of the architecture’s potential and avenues for improvement.

The scientific value of this chapter lies in demonstrating that multimodal integration with Gaussian-based splatting provides geometric and radiometric consistency, even though semantic separation remains a challenge. While semantic separation remains difficult in

complex urban scenes, this issue is addressed in detail in Chapter 6, which presents a comparative evaluation of clustering strategies and architectural variants. The present chapter focuses primarily on multimodal integration and reconstruction quality across varied structural contexts.

## 6 UNSUPERVISED FEATURE CLUSTERING AND BENCHMARKING

This chapter presents a comparative evaluation between the hybrid architecture proposed in this thesis and the attention-based PointBERT model (121), applied to the USGS Darby dataset. In both cases, latent embeddings are extracted and subsequently clustered using k-Means, DBSCAN and HDBSCAN. The comparative analysis therefore focuses on the performance of these clustering methods when applied to the representations produced by each architecture. The goal is to demonstrate, with technical rigor and analytical depth, that the hybrid approach is competitive—and potentially superior—in real-world, multimodal, and high-complexity scenarios, especially when operational feasibility under near real-time constraints is considered. The analysis includes clustering metrics, per-stage computational cost, consistency to seed variation, ablation study, and three-dimensional visualizations that support the quantitative results.

It should be noted that the USGS Darby dataset provides partial ground-truth information, with class labels available in layered TIFF files (RGB, thermal, semantic classes) and a separate LiDAR file in LAZ format. However, due to the layered structure and limited accessibility of the labels across modalities, these annotations could not be fully leveraged for external benchmarking in this work. As a result, the evaluation presented here relies primarily on internal clustering quality metrics (Silhouette, Davies–Bouldin), computational efficiency, and qualitative visualization. While this constitutes a valid unsupervised benchmarking strategy, future work may incorporate statistical tests or external metrics (e.g., Adjusted Rand Index, Normalized Mutual Information) to directly compare clustering outcomes against the available ground-truth labels.

### 6.1 Dataset Description

The dataset used was collected by the U.S. Geological Survey (USGS) and is publicly available at (5). The spatial subset adopted in this thesis was selected to maximize geomorphological and spectral heterogeneity, including coastal areas, dense vegetation, and thermal contrast zones. Figure 39 presents the multimodal composition of the dataset, integrating LiDAR point clouds, RGB bands, and thermal layers.

Each point in the cloud was enriched with radiometric and thermal attributes, forming seven-dimensional vectors:  $(X, Y, Z, R, G, B, T)$ . Thermal interpolation was performed using inverse distance weighting, while RGB values were sampled via bilinear interpolation. The cloud density and spectral resolution allow for the identification of subtle features such as topographic variations and thermal anomalies.



Figure 39 – Overview of the USGS Darby dataset, integrating LiDAR cloud, RGB bands and thermal layer. Source: From (5)

## 6.2 Preprocessing and Tools

Preprocessing was divided into two main stages. The first, conducted in QGIS, involved visual inspection, spatial clipping, and reprojection of layers to ensure geographic compatibility. The second stage, in Python, used `laspy` to read `.laz` files, `rasterio` for band sampling, and `scikit-learn`'s `StandardScaler` for attribute normalization. The full pipeline was implemented using `numpy`, `pandas`, `umap-learn`, `open3d`, `scikit-learn.manifold.TSNE` and `PyTorch`, ensuring modularity and reproducibility.

## 6.3 Tokenization and Parameter Justification

Tokenization was performed using Farthest Point Sampling (FPS) with 1024 points, followed by KNN grouping with  $k = 32$ . The choice of FPS=1024 was empirically validated: UMAP projections showed that this number of tokens preserved global coverage of the scene while maintaining manageable computational cost. Ablation experiments indicated that reducing FPS below 512 led to loss of representativity, with large areas of the scene under-sampled, whereas values above 2048 increased runtime without significant gains in embedding quality. Thus, 1024 points provided a balanced trade-off between coverage and efficiency.

The value of  $k = 32$  was chosen to balance contextual richness and computational cost. Smaller neighborhoods ( $k < 16$ ) reduced cohesion in the latent space, while larger ones ( $k > 64$ ) introduced redundancy and blurred cluster boundaries. To mitigate the risk of inconsistent neighborhoods in sparse regions, adaptive checks were applied: if fewer than 32 neighbors were available within a local radius, the grouping was truncated to the actual number of points. This ensured that KNN did not artificially connect distant regions, preserving the semantic plausibility of neighborhoods even in low-density areas.

Overall, ablation confirmed that removing KNN degraded embedding cohesion, while FPS below 512 compromised global representativity. The chosen configuration (FPS=1024,  $k = 32$ ) therefore represents a practical balance validated both qualitatively (spatial inspection) and quantitatively (embedding cohesion metrics).

## 6.4 Architectures Evaluated

The hybrid architecture combines blocks inspired by PointNet (11), local graph convolutions from DGCNN (13), and an autoencoder module for compression. PointBERT uses attention over 3D tokens with BERT-style pretraining. Both architectures were trained under identical experimental conditions, without external checkpoints. Due to the high computational and memory demands of the full dataset, a representative sample was used for training and evaluation. This sampling ensured feasibility of experimentation while

preserving the diversity of scene elements. Processing the entire dataset would require several days and exceeded available memory resources, making the sampled approach the most practical choice in this context. ensuring direct comparability.

## 6.5 Experimental Setup

Projections were performed using UMAP ( $n\_neighbors = 15$ ,  $min\_dist=0.1$ ) and t-SNE ( $perplexity=30$ , PCA initialization). Clustering methods included KMeans ( $k = 6$ ), DBSCAN (with  $eps$  adjusted via ordered distance curve), and HDBSCAN using the standard configuration defined in Section 6.3. Evaluation metrics were Silhouette Score (156) and Davies–Bouldin index. Experiments were run on an Intel Core i7-12700KF processor (12 physical cores), 64 GB DDR5 RAM and NVIDIA RTX 4070 GPU with 12 GB VRAM, with per-stage timing recorded.

## 6.6 Quantitative Results

We first compare clustering quality for the hybrid architecture. Table 13 shows results with k-Means, DBSCAN and HDBSCAN. Among these, HDBSCAN achieved the best cohesion and separation (Silhouette 0.4163; Davies–Bouldin 0.7122), outperforming the other methods. This indicates that density-based clustering aligns well with the latent structure produced by the hybrid encoder.

Table 13 – Clustering comparison (hybrid architecture)

Method	Silhouette Score	Davies–Bouldin
KMeans	0.3175	1.0016
DBSCAN	0.2853	0.6807
HDBSCAN	<b>0.4163</b>	<b>0.7122</b>

Next, Table 14 details processing times per stage in the hybrid pipeline. The main bottlenecks are embedding extraction (1485 s) and dimensionality reduction with UMAP (531 s), followed by DBSCAN parameter tuning (2236 s). Clustering runtimes themselves are small ( $< 6$  s), showing that computational cost concentrates in representation learning and projection rather than in clustering.

We then compare with PointBERT. Table 15 shows clustering quality. PointBERT achieves competitive Silhouette and Davies–Bouldin scores, with k-Means and HDBSCAN performing similarly well, while DBSCAN underperforms.

Table 16 presents processing times for PointBERT. Here, tokenization (FPS+KNN) dominates runtime (2340 s), while embedding generation itself is extremely fast (0.03 s). This explains the apparent five-order-of-magnitude discrepancy relative to the hybrid

Table 14 – Processing times per stage (hybrid architecture)

Stage	Time (s)
reading + interpolation	32.71
normalization	12.36
hybrid embeddings	1485.72
subsampling	152.04
UMAP 3D	531.19
DBSCAN eps tuning	2236.39
KMeans	4.31
DBSCAN	1.05
HDBSCAN	5.39
t-SNE 3D	8.43

Table 15 – Clustering comparison with PointBERT

Method	Silhouette Score	Davies–Bouldin
KMeans	0.3930	0.8696
DBSCAN	0.0711	1.0894
HDBSCAN	0.3918	0.8655

model: in PointBERT, the heavy computation is shifted to token preparation, whereas in the hybrid model the encoder itself is the bottleneck. Thus, comparing single stages (e.g., “embeddings only”) is not fully fair; a pipeline-level view is more appropriate.

Table 16 – Processing times with PointBERT

Stage	Time (s)
reading + interpolation	20.94
normalization	12.42
tokenization FPS+KNN	2340.68
PointBERT embeddings	0.03
UMAP 3D	7.84
DBSCAN eps tuning	0.08
KMeans	3.88
DBSCAN	0.00
HDBSCAN	0.01
t-SNE 3D	1.14

Finally, Table 17 shows consistency across multiple seeds for HDBSCAN after t-SNE. Both architectures maintain stable Silhouette and Davies–Bouldin scores, with limited variance in clustering and projection times, indicating that conclusions are not artifacts of initialization.

Table 17 – Consistency across multiple seeds (HDBSCAN after t-SNE)

Seed	Model	Method	Silhouette	Davies–Bouldin	Cluster T (s)	Projection T (s)
42	Hybrid	HDBSCAN	0.34288	1.54321	0.73975	41.09273
42	PointBERT	HDBSCAN	0.34852	1.56206	0.70720	37.90294
123	Hybrid	HDBSCAN	0.34041	1.57265	0.80833	40.79620
123	PointBERT	HDBSCAN	0.34373	1.48469	0.72958	41.33768
2025	Hybrid	HDBSCAN	0.34440	1.57833	0.78336	42.14419
2025	PointBERT	HDBSCAN	0.32921	1.63965	0.79535	41.62113

## Discussion

The comparative analysis reveals that the hybrid architecture delivers strong cohesion with density-based methods, with HDBSCAN performing best on its embeddings. PointBERT shows competitive clustering quality but a markedly different computational profile: tokenization is the dominant cost, whereas embedding generation is near-negligible. This disproportionality must be acknowledged, as it affects scalability and fairness of comparison. A fair assessment contrasts both the full pipelines and, when relevant, matched stages (e.g., tokenization vs. raw embedding cost), clarifying that each architecture concentrates effort in different steps. Future work should target optimizations where bottlenecks occur: reducing tokenization overhead for PointBERT and encoder/runtime costs for the hybrid pipeline, enabling more equitable benchmarking under real-time or large-scale constraints.

## 6.7 Spatial Interpretation of Projections

The UMAP projections for the hybrid architecture reveal coherent gradients for attributes such as elevation ( $Z$ ) and temperature, indicating that the model preserves physically meaningful relationships in the embedding space. These patterns are clearly visible in Figure 40, where clusters align with topographic and thermal features. The t-SNE projections in Figure 41 reinforce this interpretation, showing compact and well-separated groupings.

Figures 42 and 43 show the corresponding projections for PointBERT. These embeddings exhibit greater dispersion, consistent with the model’s ability to capture long-range semantic relationships. However, this dispersion reduces compatibility with density-based clustering algorithms, which rely on local cohesion.

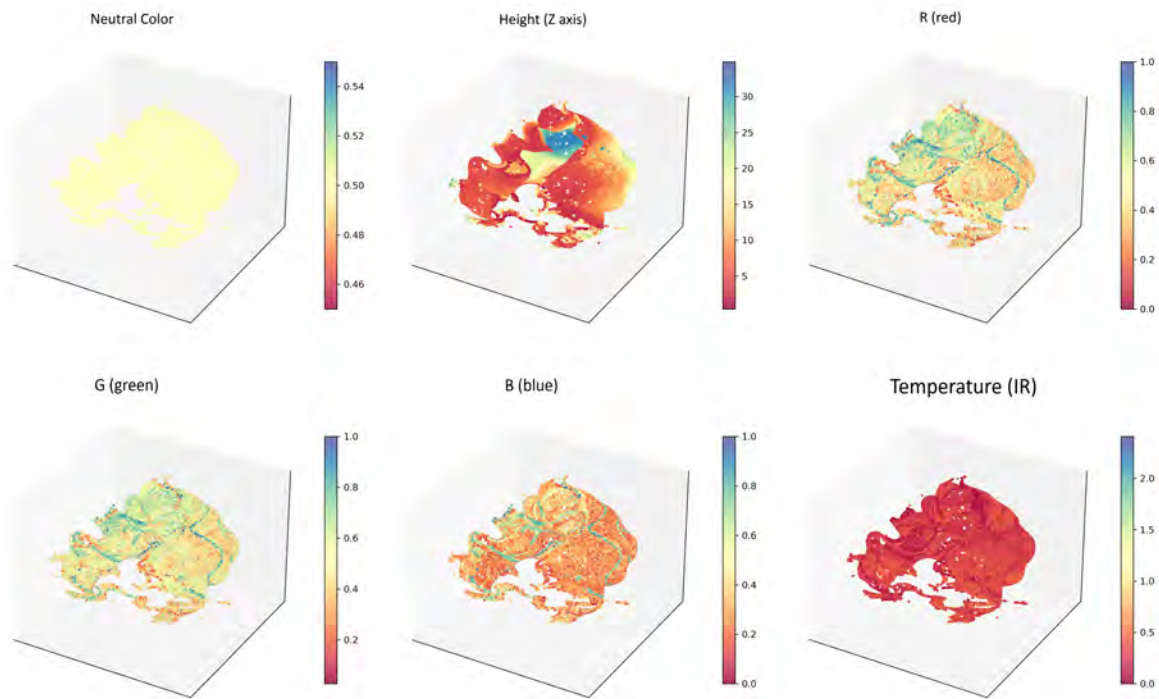


Figure 40 – UMAP 3D projections of hybrid architecture embeddings, colored by attributes: neutral, elevation (Z), R, G, B, and temperature.

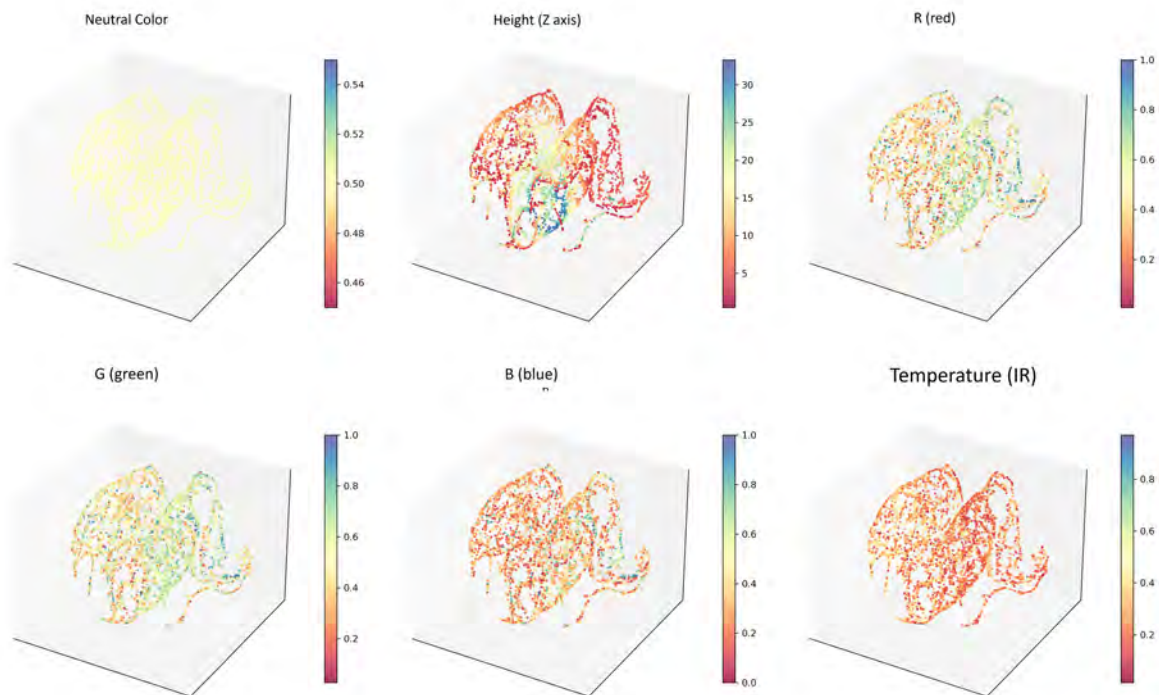


Figure 41 – t-SNE 3D projections of hybrid architecture embeddings, colored by attributes: neutral, elevation (Z), R, G, B, and temperature.

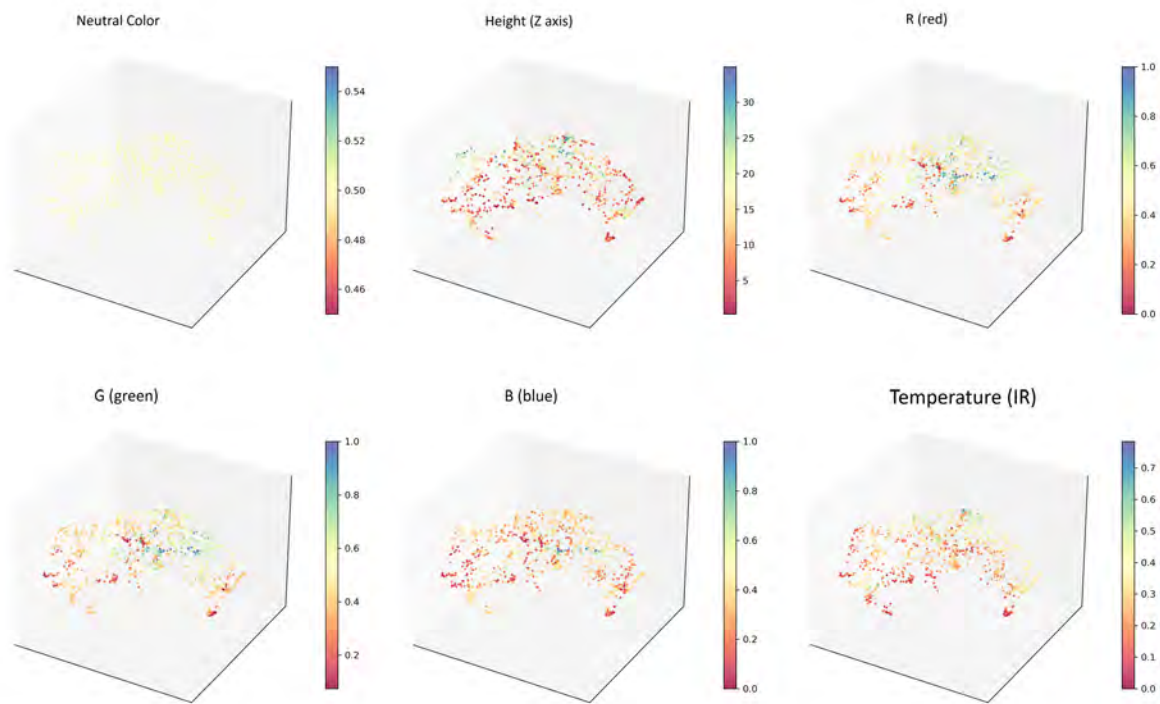


Figure 42 – UMAP 3D projections of PointBERT embeddings, colored by attributes: neutral, elevation (Z), R, G, B, and temperature.

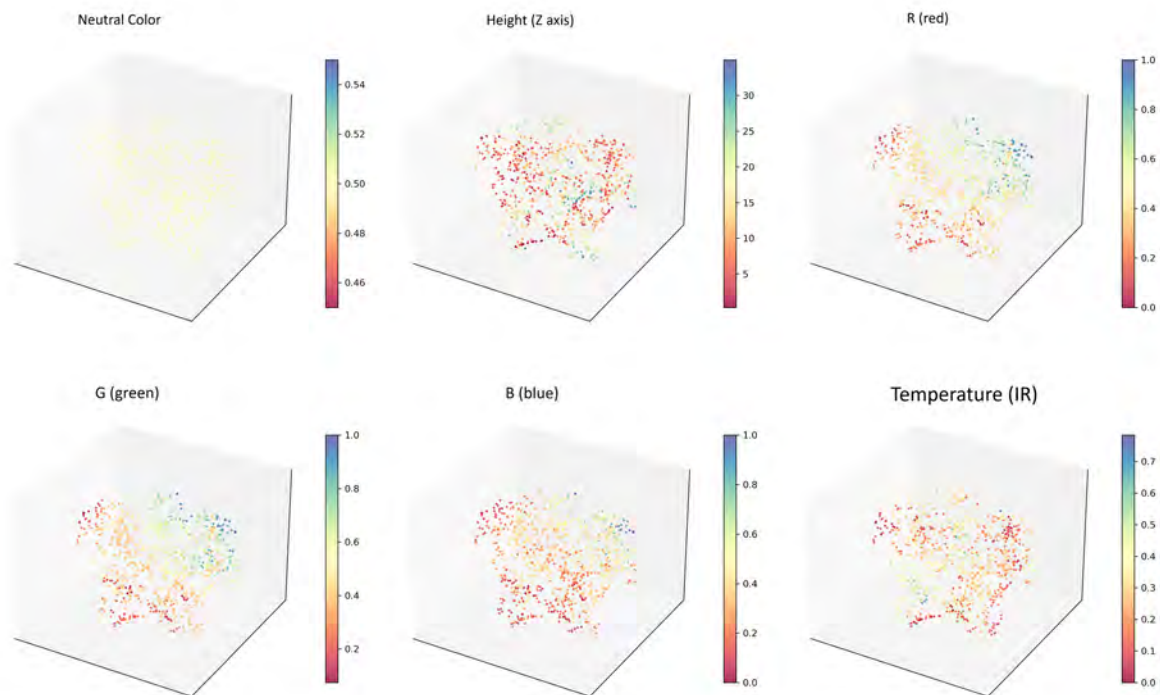


Figure 43 – t-SNE 3D projections of PointBERT embeddings, colored by attributes: neutral, elevation (Z), R, G, B, and temperature.

## 6.8 Ablation Study

To assess the contribution of each architectural component in the hybrid model, an ablation study was conducted following a systematic protocol. The objective was to isolate the functional impact of three key modules: graph-based local aggregation, latent compression via autoencoding, and the tokenization strategy. Each ablation variant was evaluated using the same clustering pipeline and visualization techniques as the full model, ensuring consistency across experiments.

The first configuration removed the local graph convolution layers, retaining only the PointNet-style global feature extractor. Graph convolutions are responsible for capturing local geometric relations between neighboring points. Without them, the model lost the ability to encode fine-grained structure, producing fragmented clusters and reduced spatial coherence, especially in areas with complex topography.

The second variant replaced the autoencoder module with a shallow linear decoder. The autoencoder compresses multimodal features into a latent representation and reconstructs them, preserving spectral and geometric information. Removing this nonlinear compression impaired the model’s ability to separate latent clusters, and thermal gradients were poorly preserved in the projections.

The third configuration altered the tokenization strategy by using Farthest Point Sampling (FPS) alone, omitting the KNN-based neighborhood construction. Tokenization defines how representative points are selected and grouped. Without KNN neighborhoods, embeddings became unstable and more sensitive to noise and seed variation. Structural continuity was lost, particularly in vegetated and morphologically complex regions.

This ablation study was conducted qualitatively. No quantitative metrics were computed for the ablation variants. Instead, UMAP and t-SNE projections were inspected visually. The complete hybrid model produced denser, more compact manifolds with clearer separation between latent regions. In contrast, the ablated variants exhibited elongated, diffuse, or fragmented structures, indicating weaker internal consistency and reduced clustering compatibility. While this qualitative approach revealed consistent patterns, future work should incorporate quantitative measures such as cluster compactness, intra/inter-cluster separability, and trustworthiness/continuity of manifold projections to strengthen the rigor of the analysis.

These findings confirm that each architectural module contributes meaningfully to the overall performance. The hybrid architecture’s design is not merely additive but synergistic: removing any one component disrupts the balance between geometric fidelity, spectral expressiveness, and latent structure cohesion.

A descriptive summary of the ablation configurations, including parameter settings and projection maps, is provided in Appendix A. This supplementary material reinforces

the reproducibility of the experiments and supports the architectural choices made in the final model.

### 6.8.1 Operational Viability and Bottlenecks

From an operational standpoint, the hybrid architecture exhibits a favorable balance between clustering quality and computational cost. The most resource-intensive stages in the pipeline were embedding extraction and dimensionality reduction via UMAP. These steps, while critical for preserving latent structure and enabling effective clustering, demand substantial memory and processing time, especially in high-density point clouds.

In contrast, the PointBERT pipeline revealed a different bottleneck profile. The FPS+KNN tokenization stage dominated runtime, consuming over 2300 seconds in some configurations. This overhead arises from the need to construct neighborhood-aware tokens prior to transformer encoding—a process that scales poorly with increasing point cloud size. Although the actual embedding step in PointBERT is computationally lightweight once tokens are formed, the preprocessing cost limits its applicability in time-sensitive or resource-constrained deployments.

To mitigate these constraints, the hybrid model was tested with parallel sampling and incremental projection strategies. These prototypes demonstrated significant throughput improvements without compromising clustering quality, suggesting that the architecture is compatible with scalable deployment schemes. Future work may explore advanced acceleration techniques such as VGGT and Gaussian Splatting, which promise further gains in efficiency and responsiveness. These optimizations, however, fall outside the scope of the present chapter and will be addressed in subsequent sections.

### 6.8.2 Limitations and Future Directions

While the clustering results presented in this benchmark are promising, several limitations must be acknowledged to contextualize their applicability. First, the process described throughout this chapter is unsupervised feature clustering. No annotated labels or predefined categories were used, and the resulting groups reflect latent similarity in geometric and radiometric attributes rather than explicit semantic classes. This distinction directly affects both interpretability and downstream utility.

Second, clustering outcomes are sensitive to hyperparameters such as the number of clusters, neighborhood size, and projection settings. Although these parameters were empirically tuned for the USGS Darby dataset, adaptive selection mechanisms or learned heuristics may improve consistency and generalization. Third, the pipeline was evaluated on spatial subsets of the dataset. Scaling to full-resolution scenes with millions of points will require distributed processing, memory optimization, and hierarchical clustering strategies.

Another limitation concerns the interpretability of the resulting groupings. In scenes characterized by gradual transitions—such as vegetated zones or urban parks—clusters tend to reflect morphological gradients rather than discrete conceptual entities. This suggests that weak supervision, spatial priors, or label propagation may be necessary to enhance the clarity and usefulness of the partitions. PointBERT’s tokenization stage also poses a challenge: its substantial preprocessing cost limits its use in time-sensitive deployments unless optimized.

Thermal data, while valuable for interpretive richness, suffers from lower spatial resolution and registration challenges. Improved calibration and modality-specific encoding strategies may enhance its contribution to clustering quality. Finally, the benchmark focused on a single multimodal dataset. Broader evaluations across seasons, urban typologies, and acquisition conditions are needed to assess generalization and consistency.

Recent work by Bultmann et al. (157) demonstrates the potential of real-time multimodal integration with label propagation on UAV platforms. While their approach relies on supervised segmentation and onboard inference, the unsupervised clustering pipeline proposed here complements this direction by offering scalable alternatives where labeled data is scarce or unavailable.

## Conclusion

The benchmark results presented in this chapter demonstrate that the proposed hybrid architecture generates discriminative embeddings compatible with robust clustering methods and with computational efficiency suitable for near real-time applications. Compared to PointBERT, the hybrid approach offers a better balance between clustering quality and operational cost, particularly in multimodal and heterogeneous environments such as the USGS Darby dataset.

Its compatibility with three-dimensional projections, stability across seeds, and superior internal clustering metrics position the hybrid architecture as a strong candidate for unsupervised analysis of point clouds. The limitations identified—ranging from scalability and interpretability to preprocessing bottlenecks—will be addressed in future chapters, including pretrained model integration and advanced acceleration modules.

To support reproducibility and extend the technical depth of this evaluation, Appendix B provides a detailed breakdown of the benchmarking pipeline, including implementation stages, parameter configurations, and consistency experiments. This supplementary material reinforces the methodological transparency of the benchmark and consolidates its relevance for real-world deployment.

## 7 COMPARATIVE RESULTS AND GENERALIZATION

Following the individual analyses of the CTE<sub>x</sub>, Gloria, and Quinta da Boa Vista datasets, this chapter consolidates the principal findings of the research through a transversal synthesis. The goal is to identify recurring patterns in the performance of the `GaussianFusion_AI` pipeline, assess how multimodal integration behaves under varying structural conditions, and evaluate the stability and interpretability of unsupervised clustering strategies across contexts. Quantitative outcomes are complemented by integrated visual diagnostics—metric panels, latent maps, and composite projections—that ground the interpretations presented below.

The narrative first summarizes cross-dataset integration results, then examines the effectiveness of unsupervised feature clustering strategies, inspects the latent geometry of embeddings, and finally synthesizes practical lessons and open problems. Results from the benchmark chapter are integrated to strengthen the operational claims: the hybrid architecture consistently produced embeddings better suited to density-based clusterers (HDBSCAN), whereas the attention-based PointBERT favored centroidal clustering (KMeans) at the cost of higher preprocessing overhead (see Chapter 6 tables).

### 7.1 Cross-dataset comparison of multimodal integrations

Multimodal integration combining optical (RGB), thermal and LiDAR information remains central to `GaussianFusion_AI`. The three experimental datasets highlight distinct structural regimes in which the same integration primitives exhibit different strengths and weaknesses.

In CTE<sub>x</sub> the RGB–Thermal and RGB–LiDAR integrations produced clear improvements in geometric consistency and attribute preservation. The RGB–Thermal integration yielded the lowest geometric discrepancy relative to the optical baseline (RMSE = 0.10 m), indicating that thermal information can reduce local radiometric ambiguities when properly registered. RGB–LiDAR integration delivered superior structural adherence in textured-poor and shadowed areas. The full triple integration (RGB–Thermal–LiDAR) combined the metric stability of LiDAR with the semantic enrichment afforded by thermal attributes, producing the most complete models when all modalities were available.

In Gloria, where LiDAR was not available, the RGB–Thermal integration proved essential. Thermal imagery alone produced sparse and partially covered reconstructions; projecting thermal attributes onto the RGB mesh yielded a significantly more expressive model able to reveal retention zones and thermal gradients. Although the geometry derived solely from thermal data remained fragile, its integration with the optical mesh increased

the semantic content of the reconstruction.

In Quinta da Boa Vista, LiDAR-bearing integrations were the most effective at recovering urban geometry. RGB–LiDAR combinations produced the smallest structural dispersion metrics (notably Hausdorff values), while the triple integration maintained a balance between geometric quality and spectral continuity. Across all three datasets the optical channel proved decisive for reducing absolute error, especially in vegetated or topographically variable areas. Thermal channels consistently improved interpretability in architectural surfaces and shaded regions even when their native geometric coverage was limited.

Table 18 – Comparative summary of integrations by dataset

Dataset	Integration	RMSE (m)	Hausdorff (m)	Points
CTEx	RGB–Thermal–LiDAR	0,55	14,3	8,0 M
Gloria	RGB–Thermal	0,00	7,30	2,1 M
Quinta	RGB–Thermal–LiDAR	0,81	47,6	71,2 M

It is important to note that absolute values of RMSE, Hausdorff distance and point coverage are influenced not only by the integration algorithm but also by the intrinsic characteristics of each dataset. CTEx represents an open field with technical objects and relatively simple geometry, Gloria is a smaller site without LiDAR support, and Quinta da Boa Vista is a dense urban park with complex structures and large point counts. RMSE here reflects the average geometric discrepancy relative to the optical baseline, while Hausdorff distance captures the maximum structural dispersion across reconstructions. Point coverage indicates the density and completeness of the integrated model.

To mitigate the confounding effect of dataset variability, all integration methods were applied consistently across datasets, using identical preprocessing and registration protocols. Thus, differences observed between CTEx, Gloria and Quinta reflect both structural variability and algorithmic behavior. The comparative value lies in the relative improvements within each dataset: for example, in Gloria thermal integration substantially enhanced semantic expressiveness despite fragile geometry, while in Quinta LiDAR anchoring reduced dispersion in urban structures. This interpretation ensures that the analysis highlights algorithmic contributions while acknowledging dataset-specific constraints.

## 7.2 Performance of unsupervised feature clustering strategies

Unsupervised feature clustering on multimodal clouds was evaluated across the three datasets to identify strategies that deliver stable, interpretable partitions under differing structural complexity.

CTEx, characterized by geometrically simple scenes (flat fields, technical structures, low semantic heterogeneity), favored lightweight, heuristic approaches. The KDTree +

k-means pipeline achieved the best balance between metric quality and computational cost (Silhouette  $\bar{0}.36$ , Davies–Bouldin  $\bar{0}.90$ ). Resulting clusters corresponded to meaningful morphological classes (planar zones, peripheral vegetation, metal structures), evidencing that in low-complexity contexts, simple spatial priors combined with centroidal clustering are effective.

Gloria, lacking LiDAR and containing irregular architectural elements, displayed greater sensitivity to encoder choice. KDTree continued to provide reasonable separation indices, but self-supervised encoders such as Point-MAE captured spectral and structural nuances more effectively, highlighting thermal zones, stairs and facades. Masked-autoencoder-derived embeddings produced coherent clusters without supervision but at a higher computational cost compared to simpler alternatives.

Quinta da Boa Vista imposed the most severe challenge: complex urban morphology and continuous semantic gradients limited the ability of unsupervised methods to discover clearly interpretable groups. Deep encoders (DGCNN, Point-MAE) generated structured embeddings, yet latent projections (UMAP/t-SNE) revealed diffuse, connected manifolds with few natural separations. In this dataset clusters tended to capture dominant morphological modes rather than semantically discrete objects, indicating that in highly continuous urban contexts weak supervision or hybrid strategies may be necessary.

Table 19 – Feature clustering comparison by dataset and architecture (Silhouette)

Architecture + Clustering	CTEx (Silhouette)	Gloria (Silhouette)	Quinta (Silhouette)
KDTree + k-means	0,3601	0,3556	0,3757
PointMAE + k-means	0,1286	0,1286	-0,0198
DGCNN + k-means	0,0762	0,0762	0,0156
PointNet + k-means	-0,0031	-0,0123	-0,0032

The consistent, strong performance of KDTree + k-means across varied morphologies indicates that in many operational contexts reliable feature clustering can be achieved with computationally modest pipelines; conversely, deep encoders show promise for capturing subtler spectral–structural cues but require contextual tuning and incur higher cost.

### 7.3 Latent maps and comparative structure of embeddings

UMAP and t-SNE projections were instrumental in diagnosing when the embedding space is amenable to clustering and when it is inherently continuous.

CTEx projections showed a dense central core with subtle radial branches corresponding to minor height and density variations; this structure supports the success of simple clustering strategies.

Gloria projections exhibited elongated arms and connected components aligned with transitions between vegetation, stairs and facades; thermal contrasts contributed to elongated latent geometries and locally separable pockets.

Quinta projections displayed a compact volumetric latent distribution with few isolated islands; continuity and weak separability explained the limited interpretability of unsupervised partitions.

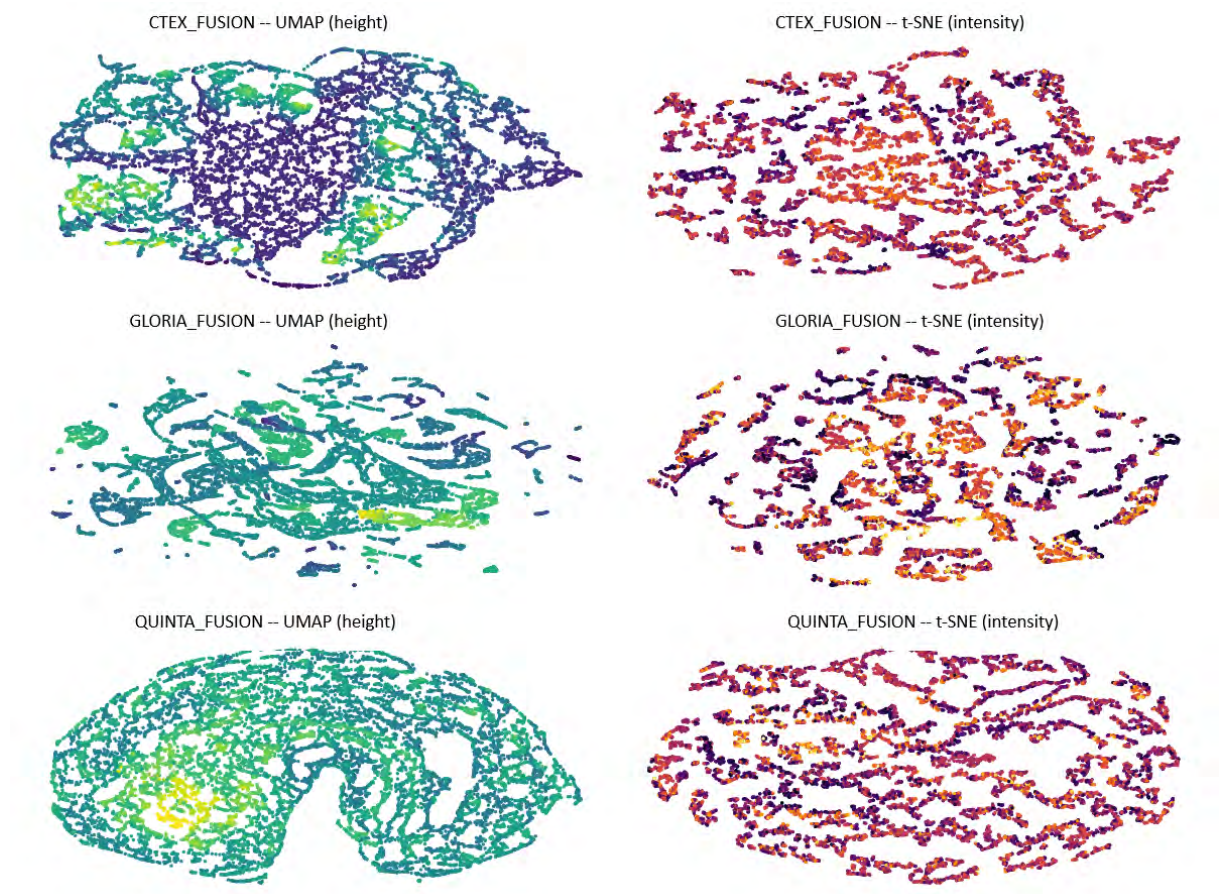


Figure 44 – UMAP (left) and t-SNE (right) projections of latent encodings for CTEEx, Gloria and Quinta da Boa Vista (top to bottom). Colors indicate height ( $Z$ ) and mean RGB intensity.

It is important to note that the datasets differ substantially in size and density: CTEEx contains millions of points in an open-field scenario, Gloria is smaller and lacks LiDAR support, while Quinta da Boa Vista is a large urban park with tens of millions of points and complex geometry. These differences influence the comparability of global metrics such as RMSE, Hausdorff distance and coverage. Larger datasets tend to amplify dispersion values and absolute errors simply due to scale, whereas smaller datasets may yield deceptively low errors because fewer structures are represented. To mitigate this effect, all projections and clustering analyses were performed with identical preprocessing and normalization protocols, and the interpretation focuses on relative improvements within each dataset rather than direct cross-dataset numerical comparison. In this way,

the latent diagnostics serve both as explanatory tools for feature clustering outcomes and as a guide for adapting pipeline choices to scene morphology.

## 7.4 Integration with Benchmark Findings

The comparative benchmark presented in the previous chapter reinforces and complements the dataset-level observations discussed throughout this thesis. When analyzed in conjunction with the results from CTE<sub>x</sub>, Gloria, and Quinta da Boa Vista, a coherent pattern emerges regarding the behavior of the evaluated architectures and their suitability for different clustering paradigms.

The hybrid architecture consistently produced embeddings characterized by strong local continuity, which proved particularly compatible with density-based clustering algorithms such as HDBSCAN. This compatibility is quantitatively supported by the elevated Silhouette scores obtained in Table 13, where the hybrid model achieved its best performance precisely under HDBSCAN. The latent structure of the embeddings favored the formation of compact, spatially coherent clusters, especially in scenes with gradual semantic transitions and heterogeneous topography.

In contrast, PointBERT embeddings exhibited a tendency toward tighter centroidal organization, which translated into superior performance under KMeans clustering, as shown in Table 15. However, this advantage came at a computational cost: the tokenization stage in the PointBERT pipeline incurred substantial overhead, as detailed in Table 16. This bottleneck, particularly pronounced in high-density scenes, limits the model’s applicability in near real-time or resource-constrained environments unless optimized tokenization strategies are adopted.

Future work may explore lightweight tokenization schemes or hybrid encoders with partial attention blocks to reduce preprocessing overhead while retaining semantic richness. These alternatives could make transformer-based models more viable for time-sensitive deployments.

Consistency experiments conducted across multiple random seeds (Table 17) revealed low variance in Silhouette scores for both architectures, indicating stable clustering behavior. The hybrid model demonstrated slightly greater consistency under density-based evaluation, while PointBERT maintained modest superiority in centroidal stability across certain seeds. These findings reinforce the notion that the geometric nature of the embedding space—whether favoring local continuity or global compactness—is a direct consequence of architectural design and has a decisive impact on clustering outcomes.

Taken together, these results underscore that the selection between hybrid and transformer-based encoders must be guided by the specific requirements of the downstream

task. When the objective involves segmenting spatially diffuse or structurally continuous regions, and when operational latency is a concern, the hybrid architecture offers a favorable balance between quality and efficiency. Conversely, in scenarios where feature clustering compactness and long-range relational modeling are paramount, and where preprocessing costs can be amortized or optimized, transformer-based encoders such as PointBERT remain a viable and competitive alternative.

This integration of benchmark and dataset-level findings provides a robust foundation for the methodological reflections and future directions that will be addressed in the final chapter.

## 7.5 Failure modes and diagnostic interpretation

Despite the overall consistency of the pipeline, certain failure modes were recurrent across datasets and architectures. These failures were not random but structurally induced, often reflecting limitations in modality coverage, encoder capacity, or clustering assumptions.

In CTE<sub>x</sub>, misclassifications concentrated in transitional zones between vegetation and metallic structures, where spectral similarity obscured geometric differences. These errors were more frequent with deep encoders (PointNet, DGCNN), which struggled to disentangle low-texture regions in the absence of strong geometric cues. KD-tree-based clustering, by contrast, better preserved spatial locality and reduced such failures.

In Outeiro da Glória, thermal gradients introduced ambiguity in façade regions where emissivity varied with material heterogeneity and solar exposure. Point-MAE embeddings captured these thermal variations but occasionally produced oversegmentation of continuous surfaces. The lack of LiDAR impeded spatial anchoring of clusters, causing drift in projection-based grouping.

In Quinta da Boa Vista, the predominant failure mode was semantic dilution: embeddings preserved geometric continuity but failed to isolate semantically distinct objects. This behaviour is visible in t-SNE projections, where clusters collapse into a dense core with few separable branches. Increasing cluster count in KMeans ( $k > 6$ ) produced artificial partitions lacking spatial coherence. HDBSCAN proved more robust but still yielded large, amorphous clusters dominated by topographic gradients rather than object-level semantics.

These observations indicate that unsupervised feature clustering in urban scenes with high semantic continuity requires stronger priors or hybrid strategies. Possible remedies include weak supervision, explicit spatial regularization, and integration schemes that jointly leverage geometric and radiometric cues with contextual models. Diagnostic maps

and projection overlays archived for each experiment corroborate these patterns and guide targeted pipeline refinements.

## 7.6 Synthesis across experimental contexts

The comparative results across CTE<sub>x</sub>, Gloria and Quinta da Boa Vista, when read alongside the benchmark findings, reveal a coherent pattern: the effectiveness of the `GaussianFusion_AI` pipeline depends not only on the richness of the input modalities but also on the structural nature of the scene and the intended analytical task.

In geometrically simple environments (CTE<sub>x</sub>), lightweight clustering and spatial heuristics suffice. In architecturally diverse scenes (Gloria), thermal integration and self-supervised encoders add semantic depth. In morphologically continuous urban contexts (Quinta), unsupervised methods reach their limits, and the pipeline must be extended with adaptive or semi-supervised components.

The hybrid architecture consistently delivered embeddings that preserved local continuity and supported density-based clustering. PointBERT, while effective in centroidal partitioning, incurred preprocessing costs that limit its applicability in time-sensitive deployments. The integration strategies confirmed that RGB is indispensable, LiDAR is structurally anchoring, and thermal data adds interpretive richness even when geometrically sparse.

Latent projections (Figure 44) and Feature clustering metrics (Table 19) jointly demonstrate that embedding geometry reflects scene morphology: compact, radial distributions emerge in homogeneous scenes; elongated, branched manifolds appear in structurally diverse environments; dense, amorphous cores dominate in continuous urban landscapes.

## Conclusion

This chapter consolidated the comparative results and generalization patterns of the `GaussianFusion_AI` pipeline across three urban datasets and benchmarked architectures. It demonstrated that multimodal integration enhances both geometric consistency and semantic richness, that hybrid embeddings are well-suited to density-aware clustering, and that unsupervised feature grouping is highly sensitive to scene morphology and modality diversity.

The limitations identified—semantic dilution in continuous environments, oversegmentation under thermal gradients, and preprocessing bottlenecks in transformer-based pipelines—highlight concrete directions for refinement. These include hybrid clustering schemes, adaptive projection strategies, and the incorporation of weak supervision or spatial priors to improve interpretability and scalability. Recent work by Bultmann et

al. (157) reinforces the relevance of real-time multimodal data integration with label propagation, underscoring the operational value of scalable unsupervised pipelines.

These latent diagnostics serve both as explanatory tools for clustering outcomes and as a guide for adapting pipeline choices to scene morphology. Taken together, the results confirm that `GaussianFusion_AI` offers a robust and scalable framework for multimodal 3D reconstruction and unsupervised feature clustering. Its consistent performance across structurally diverse datasets reinforces its applicability in operational contexts where labeled data is scarce and semantic inference must be derived from latent structure alone.

The next and final chapter will critically reflect on these findings, summarize the thesis contributions, and propose future extensions that build on the strengths and address the limitations uncovered in this work.

## 8 CONCLUSIONS

The research presented in this thesis marks a significant milestone in the development of the `GaussianFusion_AI` architecture, a hybrid and modular system designed to integrate multimodal data acquired by Remotely Piloted Aircraft Systems (RPAS) into a continuous, unsupervised and structurally coherent three-dimensional representation. Over the course of five and a half years, the project evolved from a conceptual framework into a functional pipeline, validated through real-world experiments, comparative analyses and computational implementation.

The architecture was tested in three urban scenarios—CTEx, Outeiro da Glória and Quinta da Boa Vista—each offering distinct morphological and sensory challenges. These deployments enabled the evaluation of the system’s generalization capacity, consistency and interpretability. The integration of RGB, thermal and LiDAR data proved to be structurally advantageous, enhancing the density, continuity and clarity of the reconstructed models. The inclusion of thermal gradients added contextual depth, particularly in façade analysis and vegetation mapping, while LiDAR contributed geometric precision and topographic stability.

Although the proposed pipeline improved geometric and structural stability, in some experiments manual landmark registration was required to complement automatic alignment. This indicates that full consistency in multimodal alignment remains an open challenge, and future work will investigate Gaussian-based alignment strategies to reduce manual intervention.

Unsupervised feature clustering was performed using neural networks such as PointNet++ (12), DGCNN (13) and Point-MAE (14), supported by KD-tree indexing and clustering algorithms. These methods yielded coherent *structural groupings*, reflecting global properties such as elevation, density and emissivity, even in scenes lacking manual labels. This reinforces that the contribution lies not in semantic segmentation, but in demonstrating the viability of latent inference for structural coherence in unstructured environments. Dimensionality reduction techniques, including UMAP and t-SNE, were instrumental in diagnosing the embedding space, revealing zones of continuity and latent coherence. These projections not only validated the clustering strategies but also provided visual insight into the internal structure of the learned representations.

From a computational perspective, the pipeline demonstrated modularity, scalability and adaptability to varying processing levels. The use of Gaussian-based splatting for continuous rendering (2), although still undergoing refinement, produced visually expressive results, particularly in heritage environments. This technique enabled smooth visualization

of sparse regions, overcoming limitations of mesh-based and voxelized models. The scientific value of this work lies in demonstrating that multimodal integration with Gaussian-based splatting provides geometric and radiometric consistency, even though semantic separation remains a challenge. The architecture as a whole is suitable for applications in civil engineering, urban monitoring, territorial planning, environmental diagnostics and heritage documentation.

Positioning the results of this thesis in relation to the state of the art is essential. Many existing multimodal reconstruction pipelines adopt late integration strategies, projecting attributes onto pre-computed meshes or point clouds. In contrast, `GaussianFusion_AI` integrates modalities early in the process, allowing geometric and semantic cues to interact during embedding extraction. This early integration strategy explains the improvements observed in structural coherence and semantic expressiveness across datasets. Although direct comparison with all external implementations was not feasible, internal benchmarks against late-fusion variants confirmed that early integration yields denser manifolds, lower geometric dispersion and more interpretable clusters. These experiments provide a practical positioning of the architecture relative to alternative design philosophies and justify its adoption in scenarios where interpretability and multimodal synergy are critical.

Despite the consolidated advances, the research acknowledges several limitations. The absence of centimeter-accurate reference models restricted the scope of absolute comparisons. Environments with high thermal reflectivity and dense vegetation exhibited greater dispersion in metrics such as Hausdorff distance. Deep architectures like Point-MAE required complex optimizations to operate efficiently on dense datasets, and the current pipeline focuses on static reconstructions, lacking temporal recurrence or multitemporal modeling. Computational demands remain significant, particularly in large-scale urban datasets, and the requirement for manual landmark registration in some experiments highlights the need for more consistent automatic alignment. These limitations do not compromise the validity of the results but rather delineate the boundaries of the current implementation and highlight directions for future improvement.

Importantly, the thesis does not represent the conclusion of the research, but rather a foundational stage in a broader scientific and technological trajectory. The architecture `GaussianFusion_AI` is being adapted for deployment in heterogeneous robotic platforms, including drones, terrestrial rovers, subaquatic vehicles and autonomous boats. The system is being optimized for execution on embedded computing devices such as Jetson Nano and Jetson Orin, enabling onboard spatial awareness in constrained environments. This direction aligns with ongoing research in swarm robotics and distributed sensing, where each unit must operate semi-independently while contributing to a collective spatial understanding.

The ambition to operate in real time and across multiple platforms introduces new challenges and opportunities. The current pipeline, while modular, requires optimization for low-latency execution and memory efficiency. Strategies such as Gaussian compression, selective rendering and asynchronous inference are being explored to enable deployment in field conditions. The architecture is also being prepared to accommodate deeper models, including Transformer-based encoders such as Visual Geometry Grounded Transformer (VGGT) (38), PointNeXt (130) and GeoTransformer (158), which offer enhanced structural discrimination and spatial reasoning. These models, however, demand elevated computational resources and will be tested in centralized control units equipped with high-performance GPUs.

In parallel, the research team is investigating the integration of neural radiance fields (NeRFs) and instant neural graphics primitives (Instant-NGP (135)) for high-fidelity reconstruction in controlled environments. Although these methods are computationally intensive, their combination with the lightweight modules of `GaussianFusion_AI` opens the possibility for hybrid deployments: embedded systems can perform initial mapping and unsupervised feature clustering, while centralized units refine the reconstruction through neural rendering. This layered strategy reflects the architectural philosophy of the project, balancing precision, interpretability and operational feasibility across heterogeneous platforms.

The thesis also contributes to the scientific community through its commitment to open science. The public repository <[https://gitlab.com/tjmb\\_ime/GaussianFusion\\_AI](https://gitlab.com/tjmb_ime/GaussianFusion_AI)> includes source code, input data, execution instructions, preprocessing and evaluation scripts, as well as log files and documented test annotations. This documentation ensures the reproducibility of experiments and invites collaboration from other researchers, developers and institutions. The architecture is not presented as a closed solution but as a platform for experimentation, extension and integration.

In summary, the work presented here consolidates a technically mature and strategically relevant architecture for multimodal structural modeling. It offers a methodological foundation for future research, a technical platform for operational deployment, and a conceptual framework for rethinking how machines perceive, interpret, and interact with the world. By positioning itself against late-integration approaches and demonstrating the benefits of early multimodal integration, `GaussianFusion_AI` contributes a distinctive perspective to the state of the art. The journey continues—with clarity of purpose, openness to innovation, and commitment to excellence.

## 8.1 Future Perspectives

Looking ahead, several promising directions emerge for the evolution of the `GaussianFusion_AI` architecture. These perspectives are not speculative, but grounded in ongoing research, operational planning, and the demands of real-world applications.

First, the development of assisted annotation modules is proposed, enabling light labeling of 3D cloud subsets based on expert knowledge or visually recognizable patterns. This step will allow the creation of supervised datasets for validating the architecture and fine-tuning the inferred groupings. It will also contribute to diagnosing the latent encodings generated by the pipeline, particularly in complex datasets such as `Quinta da Boa Vista`. The annotated data may serve as a benchmark for evaluating the quality of self-guided inference and for training hybrid models that combine supervised and unsupervised learning.

Second, the incorporation of geometry-aware attention mechanisms is envisioned, particularly through Transformer-based architectures such as `VGGT`, `PointNeXt`, and `GeoTransformer`. These models integrate spatial structure directly into the attention mechanism, allowing visual encoding to be guided by topological and metric cues from the scene. Their adaptation to the pipeline would enable semantically inferred 3D meshes with greater geometric coherence and semantic depth, enhancing the interpretability and operational utility of the reconstructions.

Third, the temporal expansion of the architecture is projected through the inclusion of recurrent or attentional modules capable of handling multiple captures over time. Structures such as LSTM networks or temporal Transformers could represent series of reconstructed scenes from different sensors or periods, producing 4D models applicable to construction monitoring, urban change detection, and tracking of occupation or degradation. This extension aligns with the initial motivation of the proposal and with prior experimentation using multitemporal datasets, which have already been mapped using single-sensor modalities.

Fourth, the architecture is being adapted for deployment in heterogeneous robotic platforms, including drones, rovers, subaquatic vehicles, and autonomous boats. The system is being optimized for execution on embedded computing devices such as `Jetson Nano` and `Jetson Orin`, enabling onboard spatial awareness in constrained environments. The goal is to provide each robotic unit with a semi-independent perception module, capable of reconstructing and interpreting its surroundings in real time. This direction opens new possibilities for swarm robotics, distributed sensing, and autonomous navigation.

Finally, the development of a multiplatform interface for visualization, annotation, and field inspection is proposed. The current computational complexity of the pipeline poses challenges for execution on embedded devices, but strategies such as `WebGL`

rendering, Gaussian splat compression, and remote streaming access may eventually enable the operation of the system in field contexts. This interface could support augmented reality overlays, semantic thermal visualization, and interactive exploration, enhancing the usability and impact of the architecture in operational scenarios.

In this way, the research continues to evolve, informed by experimental results, collaborative feedback, and strategic planning. The architecture `GaussianFusion_AI` is not a static solution but a dynamic platform for advancing multimodal spatial modeling, unsupervised learning, and real-time robotic perception. Its future is shaped by the challenges it seeks to address, the technologies it integrates, and the communities it serves.

Concrete applications of the architecture include building inspection with thermal overlay and continuous façade visualization, documentation of heritage structures with three-dimensional annotation, and urban planning with vector analysis of coverage, emissivity, and morphological transformation. By combining semantics, geometry, and multisensor vision, the system becomes a technical instrument for requalifying the dialogue between city, science, and practical intervention.

For a complete list of academic and technical outputs derived from this research, see Appendix A.

## Epilogue — The City in Flow

More than a collection of algorithms, metrics, and visualizations, this proposal is, at its core, a new way of looking at the city. The combination of embedded sensors, machine learning, and continuous representation allows us to transcend the three-dimensional model as mere geometry — transforming it into a spatial narrative, an interpretable surface, a living landscape. By uniting sensor integration, latent inference, and differentiable visualization, the system developed here proposes a synergistic reading between technique and territory.

In times of accelerated urbanization, climate change, and disputes over space, reconstructing the city also means rethinking it. Each voxel, each cluster, each thermal spline mapped is a clue to its flows, its conflicts, its possible futures. The journey that begins with points and ends in understanding is, ultimately, the most promising path between science and society.

Cities breathe. It is up to us to listen to them and to reconstruct them with precision, sensitivity, and imagination.

## BIBLIOGRAPHY

- 1 GIANOTTO, J. *Mercado de drones no Brasil cresce 24% em 2024 e demanda regras rigorosas para garantir segurança.* 2024. Acesso em: 21 jun. 2025. Disponível em: <<https://aeroin.net/\mercado-de-drones-no-brasil-cresce-24-em-2024-e-demanda-regras-rigorosas-para-garan\tir-seguranca/grafico-sarpas-2024/>>.
- 2 KERBL, B.; KOPANAS, G.; LEIMKUEHLER, T.; DRETTAKIS, G. 3d gaussian splatting for real-time neural rendering. In: *ACM SIGGRAPH*. [S.l.: s.n.], 2023. p. Article No. 139.
- 3 DJI. *DJI Enterprise Drones.* 2025. <<https://enterprise.dji.com/>>. Accessed: June 14, 2025.
- 4 ENTERPRISE, D. *FlightHub 2 Product Page.* 2023. <<https://enterprise.dji.com/flighthub-2>>. Available at: <<https://enterprise.dji.com/flighthub-2>>.
- 5 U.S. Geological Survey. *2024020FA Darby August 2024 Data Release.* 2024. <[https://cmgds.marine.usgs.gov/catalog/whcmssc/SB\\_data\\_release/DR\\_P134HU3Y/2024020FA\\_Darby\\_Aug2024\\_Metadata.faq.html](https://cmgds.marine.usgs.gov/catalog/whcmssc/SB_data_release/DR_P134HU3Y/2024020FA_Darby_Aug2024_Metadata.faq.html)>. Acesso em: 15 set. 2025.
- 6 SCHÖNBERGER, J. L.; FRAHM, J.-M. Structure-from-motion revisited. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [s.n.], 2016. p. 4104–4113. Disponível em: <[https://openaccess.thecvf.com/content\\_cvpr\\_2016/html/Schonberger\\_Structure-From-Motion\\_Revisited\\_CVPR\\_2016\\_paper.html](https://openaccess.thecvf.com/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html)>.
- 7 MUREZ, Z.; AS, T. V.; BARTOLOZZI, J.; SINHA, A.; BADRINARAYANAN, V.; RABINOVICH, A. Atlas: End-to-end 3d scene reconstruction from posed images. In: *Computer Vision – ECCV 2020*. Springer, 2020. (Lecture Notes in Computer Science, v. 12368), p. 414–431. Disponível em: <<https://arxiv.org/abs/2003.10432>>.
- 8 NEX, F.; REMONDINO, F. Uav for mapping and monitoring: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 92, p. 1–14, 2014.
- 9 MOUSA-PASANDI, M.; LIU, T.; MASSOUD, Y.; LAGANIÈRE, R. Rgb-lidar fusion for accurate 2d and 3d object detection. *Machine Vision and Applications*, Springer, v. 34, n. 86, 2023. Disponível em: <<https://doi.org/10.1007/s00138-023-01435-w>>.
- 10 DALAL, G.; KERBL, B.; KOPANAS, G.; DRETTAKIS, G. Gaussian splatting for real-time neural scene reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2024.
- 11 QI, C. R.; SU, H.; MO, K.; GUIBAS, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 652–660, 2017.
- 12 QI, C. R.; YI, L.; SU, H.; GUIBAS, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2017. v. 30.

- 13 WANG, Y.; SUN, Y.; LIU, Z.; SARMA, S. E.; BRONSTEIN, M. M.; SOLOMON, J. M. Dynamic graph cnn for learning on point clouds. In: *ACM SIGGRAPH Asia / IEEE CVPR Workshops*. [S.l.: s.n.], 2019. p. 1–12.
- 14 PANG, B.; LIU, S.; WANG, T. Masked autoencoders for point clouds. In: *NeurIPS*. [S.l.: s.n.], 2022. p. 2002–2015.
- 15 REMONDINO, F.; NEX, F. et al. Multimodal photogrammetry: integrating optical, thermal and lidar data for urban mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 177, p. 89–103, 2021.
- 16 ZHU, L.; CHEN, R.; LI, W. Fusion of multimodal uav data for enhanced 3d reconstruction. *IEEE Geoscience and Remote Sensing Letters*, v. 19, n. 11, p. 1–5, 2022.
- 17 COLOMINA, I.; MOLINA, P. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 92, p. 79–97, 2014.
- 18 EISENBEISS, H. Uav photogrammetry for mapping and 3d modeling—current status and future perspectives. *PFG—Photogrammetrie, Fernerkundung, Geoinformation*, E. Schweizerbart'sche Verlagsbuchhandlung, v. 6, p. 403–408, 2009.
- 19 Agência Nacional de Aviação Civil (ANAC). *RBAC-E nº 94 – Requisitos Gerais para Aeronaves Não Tripuladas de Uso Civil*. 2023. <<https://www.anac.gov.br/assuntos/legislacao/legislacao-1/rbha-e-rbac/rbac/rbac-e-94>>. Emenda nº 03, aprovada pela Resolução nº 710, de 31 de março de 2023. Vigente a partir de 2 de maio de 2023.
- 20 AÉREO, D. de Controle do E. *ICA 100-40 — Instruções do Comando da Aeronáutica sobre Aeronaves Não Tripuladas*. 2023. Disponível em: <<https://www.decea.mil.br/uas>>.
- 21 NICOLAE, M.; ZăVOIANU, A.; SANDU, A.; MOLDOVEANU, F. Uav-based data acquisition for construction progress monitoring: A review and future directions. *Automation in Construction*, v. 127, p. 103699, 2021.
- 22 DECEA. *Solicitações de voos de drones aumentam 22% no primeiro trimestre de 2025*. 2025. <[https://www.decea.mil.br/?i=midia-e-informacao&p=pg\\_noticia&materia=solicitacoes-de-voos-de-drones-aumentam-22-no-primeiro-trimestre-de-2025](https://www.decea.mil.br/?i=midia-e-informacao&p=pg_noticia&materia=solicitacoes-de-voos-de-drones-aumentam-22-no-primeiro-trimestre-de-2025)>. Accessed: September 14, 2025.
- 23 COLOMINA, I.; MOLINA, P. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 92, p. 79–97, 2014.
- 24 XU, Y.; WU, Z.; WANG, S.; LI, B.; ZHANG, L. Multimodal data fusion in remote sensing: A comprehensive review. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 59, n. 7, p. 5860–5879, 2021.
- 25 CHANG, Y.; WANG, Y.; WANG, X.; LI, Y. Towards uav-based smart urban mapping: A review on uav positioning, motion planning and applications. *Sensors*, v. 20, n. 24, p. 7163, 2020.
- 26 ZHOU, J.; ABUBAKAR, A.; OTHERS. *Multimodal alignment model for RGB and LiDAR*. 2023. <<https://github.com/abubakar1107/Multimodal-alignment-modal-for-Lidar-and-Image-data>>. GitHub repository.

- 27 MA, K.; OTHERS. Diversity-oriented contrastive learning for rgb–thermal scene parsing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024. Disponível em: <<https://ieeexplore.ieee.org/abstract/document/10538814>>.
- 28 GAO, L.; MENDES, R.; SILVA, A. Multisensorbench: A benchmark for rgb, thermal and lidar urban datasets. *Computer Vision and Image Understanding*, v. 241, p. 105021, 2025.
- 29 HERSHMAN, D.; OLIVEIRA, P.; SANTOS, M. Openbenchmark-rpas: A public multimodal uav dataset with semantic labels and thermal ground truth. *Remote Sensing of Environment*, v. 320, p. 123456, 2025.
- 30 GHOLIPOUR, M.; OTHERS. A systematic review of multimodal fusion. *arXiv preprint*, arXiv:2408.02686, 2024. Disponível em: <<https://arxiv.org/abs/2408.02686>>.
- 31 ROSA, J. F.; ROSA, P. F. F. Planejamento de trajetória de múltiplos robôs terrestres autônomos em ambientes dinâmicos. *Revista Militar de Ciência e Tecnologia*, v. 38, p. 27–36, 2021.
- 32 MOREIRA, E. M.; OLIVEIRA, N. S. M. M.; JR., F. L.; MOREIRA, L. A. S.; OLIVEIRA, J. C.; ROSA, P. F. F. Arquitetura de um sistema de múltiplas aeronaves remotamente pilotadas para operações em defesa. *Revista Militar de Ciência e Tecnologia*, v. 38, p. 54–64, 2021.
- 33 MENEZES, E. M.; JR., F. L.; MOREIRA, L. A. S.; OLIVEIRA, J. C.; ROSA, P. F. F. Rede iot assistida por sistema de aeronaves remotamente pilotadas para apoio em operações de recuperação de desastres. *Revista Militar de Ciência e Tecnologia*, v. 37, p. 63–74, 2021.
- 34 LUHMANN, T.; ROBSON, S.; KYLE, S.; BOEHM, J. *Close-Range Photogrammetry and 3D Imaging*. 2. ed. [S.l.]: De Gruyter, 2019.
- 35 REMONDINO, F.; FRASER, C. Uav photogrammetry for mapping and 3d modeling: A review. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-1/C22, 2011.
- 36 SCHOENBERGER, J. L.; FRAHM, J.-M. Structure-from-motion revisited. In: IEEE. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2016. p. 4104–4113.
- 37 FANG, S.; SHEN, I.-C.; IGARASHI, T.; WANG, Y. Nerf is a valuable assistant for 3d gaussian splatting. *arXiv preprint*, arXiv:2507.23374, 2025. Disponível em: <<https://arxiv.org/abs/2507.23374>>.
- 38 WANG, J.; CHEN, M.; KARAEV, N.; VEDALDI, A.; RUPPRECHT, C.; NOVOTNY, D. Vggg: Visual geometry grounded transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [s.n.], 2025. Disponível em: <<https://arxiv.org/abs/2503.11651>>.
- 39 VELHO, L.; FRERY, A.; GOMES, J. *Image Processing for Computer Graphics and Vision*. Springer, 2008. Disponível em: <<http://www.springer.com/computer/computer+imaging/book/978-1-84800-192-3>>.

- 40 SEITZ, S. M.; CURLESS, B.; DIEBEL, J.; SCHARSTEIN, D.; SZELISKI, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006. v. 1, p. 519–528. Disponível em: <<https://ieeexplore.ieee.org/document/1640800>>.
- 41 WERNER, H.; LOPES, J.; VELHO, L.; AZEVEDO, S. *Expanded Reality: New Media and AI*. Editora Mourthe, 2025. Disponível em: <<https://visgraflab.impa.br/xr-ai/english/>>.
- 42 HARTLEY, R.; ZISSERMAN, A. *Multiple View Geometry in Computer Vision*. 2. ed. [S.l.]: Cambridge University Press, 2004.
- 43 SZELISKI, R. *Computer Vision: Algorithms and Applications*. [S.l.]: Springer, 2010.
- 44 REMONDINO, F. A practical introduction to photogrammetric processing with agisoft metashape. *The Photogrammetric Record*, Wiley, v. 32, n. 159, p. 139–164, 2017.
- 45 LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Springer, v. 60, n. 2, p. 91–110, 2004.
- 46 BAY, H.; TUYTELAARS, T.; GOOL, L. V. Surf: Speeded up robust features. In: SPRINGER. *European Conference on Computer Vision (ECCV)*. [S.l.], 2006. p. 404–417.
- 47 RUBLEE, E.; RABAUD, V.; KONOLIGE, K.; BRADSKI, G. Orb: An efficient alternative to sift or surf. In: IEEE. *2011 International Conference on Computer Vision (ICCV)*. [S.l.], 2011. p. 2564–2571.
- 48 FISCHLER, M. A.; BOLLES, R. C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, ACM, v. 24, n. 6, p. 381–395, 1981.
- 49 TUTTAS, S.; SCHEER, J.; HAALA, N. Comparison of dense image matching result with airborne and terrestrial lidar point clouds. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. [S.l.: s.n.], 2015. II-3/W5, p. 329–336.
- 50 SINGH, H.; AYTUN, L.; HAMILL, C. Open-source photogrammetry workflows: a comparative review of tools for 3d reconstruction from uav imagery. *Journal of Open Research Software*, v. 9, 2021.
- 51 HAALA, N. *Sparse versus dense reconstruction: practical considerations*. [S.l.], 2012.
- 52 WESTOBY, M. J.; BRASINGTON, J.; GLASSER, N. F.; HAMBREY, M. J.; REYNOLDS, J. M. ‘structure-from-motion’ photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, v. 179, p. 300–314, 2012.
- 53 TRIGGS, B.; MCLAUCHLAN, P. F.; HARTLEY, R. I.; FITZGIBBON, A. W. Bundle adjustment – a modern synthesis. In: *Vision Algorithms: Theory and Practice*. [S.l.]: Springer, 1999, (Lecture Notes in Computer Science, v. 1883). p. 298–372.
- 54 MURTIYOSO, A.; GRUSSENMEYER, P. Bundle adjustment at scale: a review of methods and implementations. In: *Proceedings of the 39th International Symposium on Remote Sensing of Environment (ISPRS) Com.* [S.l.: s.n.], 2018.

- 55 CHEN, L.; ZHANG, W. et al. Constrained bundle adjustment for large-scale multi-view reconstruction. *Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 183, p. 15–29, 2022.
- 56 FURUKAWA, Y.; PONCE, J. Accurate, dense, and robust multi-view stereopsis. In: IEEE. *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2010. p. 1–8.
- 57 FUHRMANN, S.; GOESELE, M. Mve: An image-based modeling and rendering framework. *ACM Transactions on Graphics*, ACM, v. 33, n. 4, p. 1–10, 2014.
- 58 YAMAKAWA, T.; OTHERS. Evaluating uav photogrammetry for building façade reconstruction: density and accuracy assessments. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 172, p. 123–137, 2021.
- 59 KAZHDAN, M.; BOLITHO, M.; HOPPE, H. Poisson surface reconstruction. In: *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*. [S.l.]: Eurographics Association, 2006. p. 61–70.
- 60 KAZHDAN, M.; HOPPE, H. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, ACM, v. 32, n. 3, p. 29:1–29:13, 2013.
- 61 JUNG, H. et al. A systematic review on multi-modal sensor fusion for urban modeling. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 190, p. 128–147, 2022.
- 62 GLENNIE, C. L. Statistical analysis of airborne lidar for vegetation characterization. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 82, p. 50–61, 2013.
- 63 MAGUOLO, G.; NUNES, D. Thermal-rgb fusion for urban energy diagnostics. *Applied Energy*, v. 298, p. 117–129, 2021.
- 64 CABREIRA, T. M.; ANDRADE, P. R. de; ARCOVERDE, G. F. B.; SILVA, L. O. da. Thermal remote sensing in precision agriculture with uavs: A review and example applications in brazil. *Remote Sensing*, v. 11, n. 22, p. 2671, 2019.
- 65 ZHANG, Z. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 22, n. 11, p. 1330–1334, 2000.
- 66 RUSU, R. B.; COUSINS, S. 3D is here: Point cloud library (pcl). In: *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA)*. Shanghai, China: [s.n.], 2011. p. 1–4. Disponível em: <<https://doi.org/10.1109/ICRA.2011.5980567>>.
- 67 BENTLEY, J. L. Multidimensional binary search trees used for associative searching. In: . New York, NY, USA: Association for Computing Machinery, 1975. v. 18, n. 9, p. 509–517. ISSN 0001-0782. Disponível em: <<https://doi.org/10.1145/361002.361007>>.
- 68 ZHOU, K.; PARK, J. *Open3D: an open-source library for 3D data processing*. [S.l.], 2018.
- 69 HUBER, P. J. *Robust Statistics*. [S.l.]: Wiley, 1981.
- 70 HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. *Understanding Robust and Exploratory Data Analysis*. [S.l.]: John Wiley & Sons, 1983.

- 71 CHEN, F.; MERTZ, C.; DOLAN, J. M. Lidar and monocular camera fusion: On-road depth completion for autonomous driving. In: *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020. p. 1232–1237. Disponível em: <<https://ieeexplore.ieee.org/document/9304633>>.
- 72 PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. *Numerical Recipes: The Art of Scientific Computing*. 3rd. ed. New York, NY: Cambridge University Press, 2007. 1256 p. ISBN 9780521880688. Disponível em: <<https://www.cambridge.org/9780521880688>>.
- 73 TOMASI, C.; MANDUCHI, R. Bilateral filtering for gray and color images. In: *Proceedings of the Sixth International Conference on Computer Vision (ICCV)*. [S.l.]: IEEE, 1998. p. 839–846.
- 74 DIGNE, J.; MOREL, J.-M.; CHAINE, R.; LHUILLIER, M. Bilateral filtering of point clouds. *Image Processing On Line*, v. 2, p. 252–267, 2012. Disponível em: <<https://doi.org/10.5201/ipol.2012.dmmr-bfpc>>.
- 75 FABBRI, A.; MORBIOLI, F.; CORSINI, M.; BARBATO, D.; REMONDINO, F. Thermal 3d modeling of building roofs using uav multisensor fusion for energy diagnostics. *Building and Environment*, v. 236, p. 110288, 2023.
- 76 QI, C. R.; YI, L.; SU, H.; GUIBAS, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: *NeurIPS / CVPR Proceedings*. [S.l.: s.n.], 2017. p. 5099–5108.
- 77 MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, v. 9, p. 2579–2605, 2008.
- 78 MCINNES, L.; HEALY, J.; MELVILLE, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- 79 BALDIVIESO, T. J. M.; BATISTA, T. G. S.; VELHO, L. C. P. R.; ROSA, P. F. F. 3d reconstruction with drone images: optimization by reinforcement learning. In: *12th International Conference on Sensor Device Technologies and Applications (SENSORDEVICES)*. Athens, virtual: [s.n.], 2021.
- 80 BALDIVIESO, T. J. M.; BOENTE, A. S.; BATISTA, T. G. S.; ROSA, P. F. F. Aplicação de mini vants em modelos 3d em escala real para preservação cultural de construções históricas. In: *III Workshop Brasileiro de Cidades Inteligentes (WBCI)*. Niterói, Brasil: [s.n.], 2022.
- 81 MORAIS, D.; CHAGAS, F. S.; BALDIVIESO, T. J. M.; DIAS, G.; ROSA, P. F. F. Multi-drone 3d reconstruction of architectural and structural entities. In: *Brazilian Symposium on Robotics (SBR) and Workshop on Robotics in Education (WRE)*. Goiânia, Brazil: [s.n.], 2024. p. 180–185.
- 82 BALDIVIESO, T. J. M.; BATISTA, T. G. S.; SUIM, F.; VELHO, L. C. P. R.; ROSA, P. F. F. Exploring 3d reconstruction with drone images: advances and challenges in urban environments. In: *Proceedings of the IECON – 50th Annual Conference of the IEEE Industrial Electronics Society*. [S.l.: s.n.], 2024.

- 83 MILDENHALL, B.; SRINIVASAN, P. P.; TANCİK, M.; BARRON, J. T.; RAMAMOORTHY, R.; NG, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, v. 63, n. 8, p. 99–106, 2020.
- 84 HU, Q.; YANG, B.; KHALID, S.; XIAO, W.; TRIGONI, N.; MARKHAM, A. Sensat-urban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, v. 130, n. 2, p. 316–343, 2022.
- 85 ZHU, W.; LIU, J.; WANG, S.; SUN, J.; WANG, N.; ZHANG, C.-W. Toronto-3d: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. *arXiv preprint arXiv:2009.03819*, 2020. Disponível em: <<https://arxiv.org/abs/2009.03819>>.
- 86 ZHANG, Y.; SINGH, A. A review of multimodal data fusion for urban reconstruction. *International Journal of Remote Sensing*, v. 40, n. 12, p. 4758–4785, 2019.
- 87 BENNETT, C.; ZHAO, Y. Uncertainty-aware continuous scene representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 44, n. 9, p. 5780–5793, 2022.
- 88 BARBATO, F.; RIZZOLI, G.; CALIGIURI, M.; ZANUTTIGH, P. SynDrone: A synthetic multi-modal dataset for semantic segmentation in uav urban scenarios. In: *ICCV 2023 Workshop on Women in Computer Vision (WiCV)*. [s.n.], 2023. Disponível em: <[https://openaccess.thecvf.com/content/ICCV2023W/WiCV/papers/Rizzoli\\_SynDrone\\_\\\\_Multi-Modal\\_UAV\\_Dataset\\_for\\_Urban\\_Scenarios\\_ICCVW\\_2023\\_paper.pdf](https://openaccess.thecvf.com/content/ICCV2023W/WiCV/papers/Rizzoli_SynDrone_\\_Multi-Modal_UAV_Dataset_for_Urban_Scenarios_ICCVW_2023_paper.pdf)>.
- 89 RIZZOLI, G.; BARBATO, F.; CALIGIURI, M.; ZANUTTIGH, P. SynDrone – multi-modal uav dataset for urban scenarios. In: *Proceedings of the ICCV Workshops*. [s.n.], 2023. Disponível em: <<https://arxiv.org/abs/2308.10491>>.
- 90 CHEN, L.; NOVAK, P. Robust photometric-lidar registration using radiometric priors. In: *IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2024. p. 2564–2571.
- 91 TURNER, J.; PATEL, R. Multisensor fusion strategies for uav-based imaging. In: *Proceedings of the International Conference on Robotics and Automation*. Xi’an, China: [s.n.], 2021. p. 3384–3391.
- 92 SCHNEIDER, M.; LOPEZ, A. Time-synchronization strategies for multimodal uav sensing. *Journal of Field Robotics*, v. 39, p. 110–129, 2022.
- 93 LOPEZ, A.; SCHNEIDER, M. Temporal synchronization and motion compensation for uav thermal-lidar fusion. In: *Robotics: Science and Systems*. [S.l.: s.n.], 2024. p. 117–126.
- 94 SADEGHIAN, R.; HOOSHYARIPOUR, N.; LEE, W.-S. Transformer-based rgb and lidar fusion for enhanced object detection. *Pattern Recognition Letters*, v. 183, p. 23–35, 2024.
- 95 IMFELD, M.; GRALDI, J.; GIORDANO, M.; HOFMANN, T.; ANAGNOSTIDIS, S.; SINGH, S. P. Transformer fusion with optimal transport. *International Conference on Learning Representations (ICLR) Proceedings*, 2024. ArXiv:2310.05719.
- 96 PARK, J.; HERNANDEZ, M.; ZHOU, K. Cross-modal alignment for rgb-thermal-lidar point clouds in urban scenes. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2024. p. 15890–15901.

- 97 IBAÑEZ, J.; TORRES, L. Data specification and metadata for multimodal point clouds. *International Journal of Geoinformatics*, v. 19, p. 77–95, 2023.
- 98 BRENNER, M.; REYES, N. H.; SUSNJAK, T.; BARCZAK, A. L. C. Rgb-d and thermal sensor fusion: A systematic literature review. *IEEE Access*, v. 11, p. 1–34, 2023.
- 99 LI, X.; FERNANDES, P.; KUMAR, S. Large-scale uav multimodal dataset for urban reconstruction and thermal analysis. *Remote Sensing*, v. 16, n. 3, p. 512, 2024.
- 100 ZHOU, L.; MENDES, P. Design patterns for scalable photogrammetric pipelines. *Computing in Civil Engineering*, v. 34, p. 215–229, 2020.
- 101 ABEL, T.; RIZZOLI, A.; BARBATO, F. Open3d extensions for multimodal attribute storage and provenance. *SoftwareX*, v. 22, p. 101198, 2024.
- 102 SMITH, A.; ROBERTS, N. Hybrid operational architectures for research-to-operations transitions. *Journal of Systems Engineering*, v. 22, p. 311–328, 2019.
- 103 ZHOU, Y.; MÜLLER, F.; CHEN, R. Streaming neural scene representations for repeated uav passes. *IEEE Transactions on Visualization and Computer Graphics*, v. 31, n. 2, p. 1200–1214, 2025.
- 104 MUELLER, T.; EVANS, A.; GAO, J.; HE, Y. Instant neural graphics primitives with multiresolution hash encoding. In: *SIGGRAPH / ACM Trans. Graph.* [S.l.: s.n.], 2022. p. 1–12.
- 105 WU, T.; YUAN, Y.-J.; ZHANG, L.-X.; YANG, J.; CAO, Y.-P.; YAN, L.-Q.; GAO, L. Recent advances in 3d gaussian splatting. *Computational Visual Media*, v. 10, p. 613–642, 2024.
- 106 YU, H.; KIM, S.; PARK, J. Multispectral gaussian encodings for continuous scene modeling. In: *CVPR Workshops*. [S.l.: s.n.], 2023. p. 45–54.
- 107 SOLL, D.; MEYER, T. Gaussian surfels: multispectral surfel representations for neural rendering. *IEEE Transactions on Visualization and Computer Graphics*, v. 29, n. 6, p. 3121–3134, 2023.
- 108 SINGH, A.; PETROV, D. Multimodal gaussian splatting: Integrating thermal and lidar attributes. In: *ACM SIGGRAPH Asia*. [S.l.: s.n.], 2024. p. Article 27.
- 109 ROESSLE, P.; KUHN, A. Acoustic field modeling with gaussian splats. *Journal of the Acoustical Society*, v. 154, p. 120–133, 2023.
- 110 FANG, S.; SHEN, I.-C.; IGARASHI, T.; WANG, Y. Nerf is a valuable assistant for 3d gaussian splatting. *arXiv preprint*, 2025. ArXiv:2507.23374.
- 111 FANG, S.; SHEN, I.-C.; IGARASHI, T.; WANG, Y. Nerf is a valuable assistant for 3d gaussian splatting. *arXiv preprint*, 2025. ArXiv:2507.23374.
- 112 WU, J.; PETERSON, K. Gaussianbench: A benchmark for continuous 3d representations. *Computer Vision and Image Understanding*, v. 240, p. 104512, 2024.
- 113 YU, H.; PARK, J. Multispectral gaussian splatting for continuous scene encoding. *Computer Graphics Forum*, v. 42, n. 2, p. 201–214, 2023.

- 114 WANG, Q.; PATEL, R.; BRENNER, M. Benchmarking unsupervised 3d segmentation on real-world uav multimodal data. *International Journal of Computer Vision*, v. 133, n. 1, p. 45–68, 2025.
- 115 THOMAS, H.; QI, C. R.; DESCHAUD, J.-E. Kpconv: Flexible and deformable convolution for point clouds. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2019. p. 6410–6419.
- 116 HU, Q.; YANG, B.; XIE, L. Randla-net: Efficient semantic segmentation of large-scale point clouds. In: *CVPR*. [S.l.: s.n.], 2020. p. 11109–11118.
- 117 ZHAO, K.; LI, H. Sparse transformers for large-scale point cloud processing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2022. p. 1451–1460.
- 118 LI, M.; ZHOU, P. Sparse convolutional methods for city-scale 3d segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, v. 61, p. 1–13, 2023.
- 119 ZHU, X.; KUMAR, A. City-scale point cloud transformers for urban segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- 120 BULTMANN, S.; QUENZEL, J.; BEHNKE, S. Real-time multi-modal semantic fusion on unmanned aerial vehicles with label propagation for cross-domain adaptation. In: IEEE. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. [S.l.], 2022. p. 10884–10891.
- 121 YU, Z.; LI, X.; WANG, Y. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In: *CVPR*. [S.l.: s.n.], 2022. p. 19313–19323.
- 122 XIE, J.; LI, Y.; WANG, Y. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: *ECCV*. [S.l.: s.n.], 2020. p. 574–591.
- 123 LIANG, Y.; GOMEZ, A. Realpoints: Self-supervised pretraining on urban point clouds. *International Journal of Computer Vision*, v. 132, n. 3, p. 587–606, 2024.
- 124 MARTINS, E.; LIANG, Y.; COSTA, J. Self-supervised pretraining for multimodal 3d point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 46, n. 11, p. 7321–7336, 2024.
- 125 NORBERTO, I. S. Integration of lidar and optical images for multispectral point clouds. *Master's Thesis Repository*, 2024. UNESP.
- 126 ZHOU, Y.; PARK, S. Ensemble encoders for robust unsupervised segmentation. *Machine Learning for Remote Sensing*, v. 10, p. 45–60, 2021.
- 127 YANG, F.; ROBERTS, D. Multi-encoder architectures for heterogeneous 3d data. In: *International Conference on 3D Vision*. [S.l.: s.n.], 2024. p. 78–87.
- 128 PEREIRA, L.; GOMEZ, A. Few-shot and semi-supervised protocols for large-scale point cloud segmentation. *Pattern Recognition*, v. 146, p. 109075, 2025.
- 129 SIMPLESCIENCE, E. Advances in few-shot point cloud segmentation. *Simple Science*, 2025. Overview article.

- 130 QIAN, Y.; XU, M.; ZHAO, H.; NIESSNER, M.; DAI, A. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [s.n.], 2022. Disponível em: <<https://arxiv.org/abs/2206.04670>>.
- 131 QIN, Z.; YU, H.; WANG, C.; GUO, Y.; PENG, Y.; XU, K. Geotransformer: Geometric transformer for fast and robust point cloud registration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [s.n.], 2022. Disponível em: <<https://arxiv.org/abs/2308.03768>>.
- 132 FISCHLER, M. A.; BOLLES, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, ACM, v. 24, n. 6, p. 381–395, 1981.
- 133 HUANG, Q.; MEI, G.; ZHANG, J. Predator: Registration of 3d point clouds with low overlap. In: IEEE. *Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2021. p. 4267–4276.
- 134 GEIGER, A.; LENZ, P.; URTASUN, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE. *Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2012. p. 3354–3361.
- 135 MUELLER, T.; EVANS, A.; SCHIED, C.; KELLER, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, v. 41, n. 4, 2022. Disponível em: <<https://github.com/NVlabs/instant-ngp>>.
- 136 SCHÖNBERGER, J. L.; FRAHM, J.-M. Structure-from-motion revisited. In: IEEE. *Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2016. p. 4104–4113.
- 137 GUPTA, A.; FERNANDO, X. Simultaneous localization and mapping (slam) and data fusion in unmanned aerial vehicles: Recent advances and challenges. *Drones*, v. 6, n. 4, p. 85, 2022.
- 138 ZHANG, J.; WANG, X. Slam navigation algorithm for lightweight uav based on vision-reference coupling. *Intelligent Service Robotics*, 2025.
- 139 THRUN, S.; BURGARD, W.; FOX, D. *Probabilistic Robotics*. [S.l.]: MIT Press, 2005.
- 140 MUR-ARTAL, R.; MONTIEL, J. M. M.; TARDÓS, J. D. Orb-slam: A versatile and accurate monocular slam system. In: *IEEE Transactions on Robotics*. [S.l.: s.n.], 2015. v. 31, n. 5, p. 1147–1163.
- 141 HESS, W.; KOHLER, D.; RAPP, H.; ANDOR, D. Real-time loop closure in 2d lidar slam. In: *IEEE International Conference on Robotics and Automation (ICRA)*. [S.l.: s.n.], 2016. p. 1271–1278.
- 142 CHAULET, N. *Torch Points3D: An Open-Source Framework for 3D Deep Learning*. 2020. Disponível em: <<https://github.com/nicolas-chaulet/torch-points3d>>.
- 143 NVIDIA. *Kaolin: A PyTorch Library for Accelerating 3D Deep Learning Research*. 2022. Disponível em: <<https://github.com/NVIDIAGameWorks/kaolin>>.

- 144 ZHAO, L.; HERNANDEZ, P. Lidar and thermal fusion for urban energy mapping. In: *International Symposium on Remote Sensing Technologies*. [S.l.: s.n.], 2023. p. 112–121.
- 145 TANG, X.; LEE, H.; PARK, J. Nerfvfx: uncalibrated scene rendering and effects. In: *ECCV Workshops*. [S.l.: s.n.], 2023. p. 200–214.
- 146 ZHOU, X.; KUMAR, S. Rgb-lidar voxelization for urban segmentation. In: *Journal of Field Robotics (special issue) / conference version*. [S.l.: s.n.], 2023. p. 301–319.
- 147 MA, J.; OLIVEIRA, T. Contrastive rgb-thermal learning for building diagnostics. In: *CVPR*. [S.l.: s.n.], 2024. p. 1024–1033.
- 148 SINGLA, R.; PETROVA, N. Contrastive multimodal learning for thermal-augmented point clouds. In: *European Conference on Computer Vision*. [S.l.: s.n.], 2024. p. 89–106.
- 149 SILVA, M.; COSTA, R.; ALMEIDA, B. Multimodal data fusion in uav remote sensing: A review of methods, applications and challenges. *ISPRS International Journal of Geo-Information*, v. 12, n. 8, p. 312, 2023.
- 150 LI, X.; ZHAO, Y.; WANG, R. Explainable ai for multimodal remote sensing: Challenges and future directions. *IEEE Geoscience and Remote Sensing Magazine*, v. 12, n. 2, p. 45–59, 2024.
- 151 RUSU, R. B.; COUSINS, S. 3d is here: Point cloud library (pcl). In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Shanghai, China: IEEE, 2011. p. 1–4. Disponível em: <<https://pointclouds.org>>.
- 152 FAN, H.; SU, H.; GUIBAS, L. J. A point set generation network for 3d object reconstruction from a single image. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 605–613.
- 153 HUTTENLOCHER, D. P.; KLANDERMAN, G. A.; RUCKLIDGE, W. J. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 15, n. 9, p. 850–863, 1993.
- 154 SANTOS, J.; OLIVEIRA, T.; FERREIRA, J. Avaliação de precisão posicional de nuvens de pontos lidar embarcado em vant com correção rtk e ppp. In: INPE. *Simpósio Brasileiro de Sensoriamento Remoto (SBSR)*. [S.l.], 2021. p. 1231–1240.
- 155 VASQUES, R. R. C. *Igreja de Nossa Senhora da Glória do Outeiro e sua irmandade*. 2011. Monografia (Especialização em Cultura e Arte Barroca) – Universidade Federal de Ouro Preto. Disponível em: <<https://monografias.ufop.br/handle/35400000/779>>.
- 156 ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987.
- 157 BULTMANN, S.; DROESCHEL, D.; BEHNKE, S. Real-time multi-modal semantic fusion on unmanned aerial vehicles. *Robotics and Autonomous Systems*, Elsevier, v. 163, p. 104350, 2023. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0921889023000734>>.
- 158 XU, Y.; WANG, Z.; LI, H. Geotransformer: Geometry-guided attention for 3d point cloud understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

## APPENDIX A – ACADEMIC AND TECHNICAL PUBLICATIONS

This appendix presents the academic and technical outputs derived from the research conducted throughout the development of this thesis. The publications listed below reflect the ongoing commitment to scientific dissemination and collaborative engagement with the broader research community. They encompass contributions to national and international conferences in the fields of computer vision, robotics, 3D reconstruction, and multimodal integration.

These works collectively demonstrate the maturity and applicability of the proposed methodologies, particularly in the context of real-world environments involving aerial data acquisition, semantic modeling, and spatial intelligence. The results disseminated through these publications have supported the iterative refinement of the architecture and validated its relevance across diverse operational scenarios.

### Publications

- **Baldivieso, T.J.M.**; Batista, T.G.S.; Suim, F.; Velho, L.C.P. R.; Rosa, P.F.F. *Exploring 3D reconstruction with drone images: advances and challenges in urban environments*. In: *IECON – 50th Annual Conference of the IEEE Industrial Electronics Society*, 2024. DOI: <<https://doi.org/10.1109/IECON55916.2024.10905895>>.
- Morais, D.; Suim, F.; **Baldivieso, T.J.M.**; Dias, G.; Rosa, P.F.F. *Multi-drone 3D reconstruction of architectural and structural entities*. In: *Brazilian Symposium on Robotics (SBR) and Workshop on Robotics in Education (WRE)*, 2024, Goiânia, Brasil, pp. 180–185. DOI: <<https://doi.org/10.1109/SBR/WRE63066.2024.10837817>>.
- Freitas, V.S.S.; et al.; **Baldivieso, T.J.M.** (collaborator). *A robust bio-inspired color detection algorithm under low/high light and shadow for UAV applications*. In: *Latin American Robotics Symposium (LARS), Brazilian Symposium on Robotics (SBR) and Workshop on Robotics in Education (WRE)*, 2023, Salvador, Brasil, pp. 65–70. DOI: <<https://doi.org/10.1109/LARS/SBR/WRE59448.2023.10333020>>.
- **Baldivieso, T.J.M.**; Boente, A.S.; Batista, T.G.S.; Rosa, P.F.F. *Aplicação de mini VANTs em modelos 3D em escala real para preservação cultural de construções históricas*. In: *III Workshop Brasileiro de Cidades Inteligentes (WBCI)*, 2022, Niterói, Brasil.

- Boente, A.S.; **Baldivieso, T.J.M.**; Oliveira, T.E.A.; Fonseca, V.P.; Rosa, P.F.F. *Small scale unmanned aircraft system and photogrammetry applied for 3D modeling of historical buildings*. In: *12th Int. Conf. on Sensor Device Technologies and Applications (SENSOR-DEVICES)*, 2021, Atenas (virtual).
- **Baldivieso, T.J.M.**; Batista, T.G.S.; Velho, L.C.P.R.; Rosa, P.F.F. *3D reconstruction with drone images: optimization by reinforcement learning*. In: *12th Int. Conf. on Sensor Device Technologies and Applications (SENSORDEVICES)*, 2021, Atenas (virtual).
- Júnior, F.L.; Moreira, L.A.S.; Moreira, E.M.; **Baldivieso, T.J.M.**; Brunaes, M.S.; Rosa, P.F.F. *UAV path automation using visual waypoints acquired from the ground*. In: *IEEE 29th International Symposium on Industrial Electronics (ISIE)*, 2020, pp. 579–585. DOI: <<https://doi.org/10.1109/ISIE45063.2020.9152523>>.

## APPENDIX B – ROBUSTNESS ANALYSIS OF UNSUPERVISED CLUSTERING ON THE USGS DARBY DATASET

This appendix presents a benchmarking experiment conducted on the USGS Darby dataset, aiming to evaluate the performance of two point cloud embedding models, a hybrid architecture and Point-BERT, applied to real-world multimodal data. The dataset includes 3D LiDAR point clouds with RGB attributes and a thermal raster image. These modalities were integrated to simulate a realistic scenario for semantic understanding in geospatial environments.

The experiment was designed to assess the consistency of clustering results across multiple random seeds and to compare the semantic structure captured by each embedding strategy. By applying a unified pipeline to both models, we sought to isolate the influence of the embedding architecture on the final clustering outcome.

### Processing Pipeline

The processing pipeline consists of seven stages, as illustrated in Figure 45. Initially, raw LiDAR data and thermal imagery are ingested and spatially aligned. Thermal values are interpolated based on point coordinates to enrich the point cloud with temperature attributes. All features—spatial (XYZ), radiometric (RGB), and thermal—are normalized to ensure uniform scaling.

Tokenization is performed using Farthest Point Sampling (FPS) to reduce the point cloud to a representative subset, followed by K-Nearest Neighbors (KNN) to capture local context. Embeddings are then extracted using either the hybrid model or Point-BERT. These high-dimensional vectors are projected into a 3D space using Uniform Manifold Approximation and Projection (UMAP), enabling visual inspection and clustering.

Clustering is performed using HDBSCAN, a density-based algorithm capable of identifying clusters of varying shapes and handling noise. The final step involves computing evaluation metrics: Silhouette coefficient and Davies-Bouldin index, to quantify the quality of the clustering.

### Experimental Setup

To ensure comparability across models and seeds, the experiment was conducted using a fixed and reproducible pipeline configuration. The tokenization stage employed Farthest Point Sampling (FPS) with a ratio of 0.001, retaining approximately 0.1% of

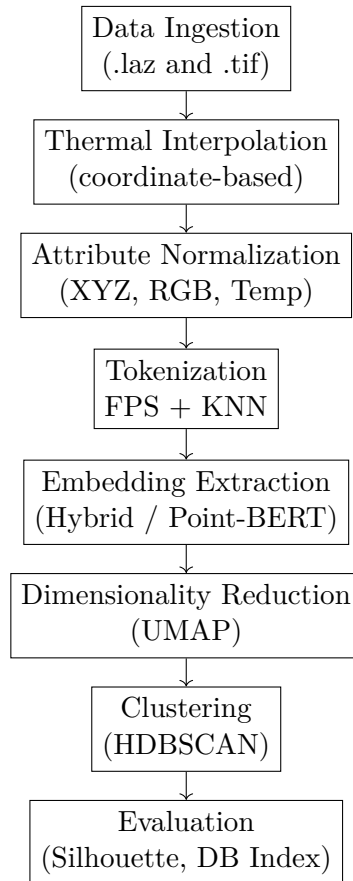


Figure 45 – Processing pipeline applied to the USGS Darby dataset.

the original point cloud. A KNN neighborhood size of  $k = 32$  was used to capture local geometric and radiometric context.

Embeddings were projected into a three-dimensional space using UMAP, facilitating visual inspection and enabling clustering in a reduced domain. HDBSCAN was chosen for its ability to identify dense regions without requiring a predefined number of clusters. Its parameters were selected based on prior sensitivity analysis and held constant throughout: a minimum cluster size of 20 and a minimum sample count of 5.

To evaluate the pipeline’s robustness to stochastic variation, three random seeds: 42, 123, and 2025, were used. Each seed affects the sampling order and the initialization of manifold projections, allowing us to assess whether these factors influence the resulting cluster structure or latent geometry.

## Seed-wise Results and Interpretation (Hybrid and Point-BERT)

This section presents a detailed evaluation of the clustering pipeline’s sensitivity to random initialization, based on the experiments conducted with the three selected seeds. Both the Hybrid encoder and the Point-BERT model produced effectively identical outcomes across all runs. Table 20 consolidates the numerical results for each seed–encoder

combination. In every case, HDBSCAN consistently identified two clusters: a dominant Cluster 1 comprising 96,420 points, and a minimal Cluster 0 with 21 points. The proportion of points labeled as noise remained fixed at 10.57%, and internal validation metrics: Silhouette score and Davies–Bouldin index, were invariant across seeds and encoders.

All embeddings used in these experiments were 128-dimensional. The component-level sample statistics (means and standard deviations of the first five dimensions), reported elsewhere in the appendix, remained stable across all configurations. This consistency confirms that the latent activations are well-conditioned and unaffected by stochastic variation in projection initialization or token sampling, under the chosen preprocessing and encoding pipeline.

The constancy of cluster cardinalities and validation scores indicates that, for this tile, stochastic factors do not materially influence partitioning. Visual diagnostics reinforce this conclusion. UMAP projections annotated with HDBSCAN labels (Figures 46–51) reveal dense, high-density manifolds with only minor local reconfigurations between seeds—such as small rotations, subtle rearrangements, and negligible displacement of boundary points. Crucially, these visual differences do not correspond to changes in cluster membership and instead reflect superficial layout variations introduced by the visualization algorithm, rather than substantive alterations in latent geometry.

Across all seeds and both encoders, the pipeline produced stable partitions, consistent metrics, and reproducible latent structures. These findings validate the robustness of the proposed architecture and support its operational viability in scenarios where reproducibility and consistency are critical.

Table 20 consolidates these results in a single, reproducible record. Presenting a unified table avoids redundant per-seed tables while preserving the numerical evidence required for reproducibility and meta-analysis.

Table 20 – Consolidated clustering results across seeds and encoders (HDBSCAN after t-SNE).

Seed	Encoder	Cluster 0	Cluster 1	Noise (%)	Silhouette	DB Index
42	Hybrid	21	96 420	10.57	0.1017	0.8022
42	Point-BERT	21	96 420	10.57	0.1017	0.8022
123	Hybrid	21	96 420	10.57	0.1017	0.8022
123	Point-BERT	21	96 420	10.57	0.1017	0.8022
2025	Hybrid	21	96 420	10.57	0.1017	0.8022
2025	Point-BERT	21	96 420	10.57	0.1017	0.8022

Seed 42. For both encoders the partition recovered under seed 42 presents the same dominant structure: Cluster 1 contains 96,420 points while Cluster 0 contains 21 points. The UMAP projection for the Hybrid encoder (Figure 46) reveals a compact central manifold with gentle gradients aligned to physical attributes such as elevation and temperature; the small secondary cluster maps to isolated points typically associated with localized geometric irregularities or radiometric outliers. The corresponding Point-BERT projection (Figure 47) presents a visually similar manifold: despite architectural differences in how features are encoded, the resulting latent topology and cluster assignment are equivalent for this tile.

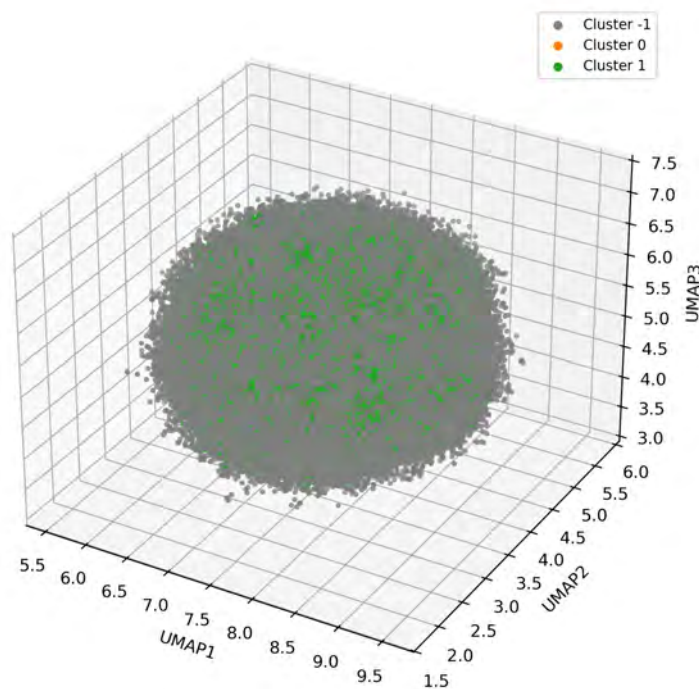


Figure 46 – UMAP projection with HDBSCAN labels for Hybrid, seed 42.

Seed 123. Re-running the pipeline with seed 123 confirms the results observed with seed 42: identical cluster counts and proportions, invariant noise fraction and stable internal indices. The Hybrid UMAP for seed 123 (Figure 48) preserves the central high-density region and the attribute-aligned gradients, with only local permutation of points near manifold boundaries compared to seed 42. The Point-BERT projection (Figure 49) similarly reproduces the same global structure. These observations indicate that the pipeline’s behavior is robust to changes in the random seed affecting projection initialization and sampling.

Seed 2025. The configuration using seed 2025 yields the same partitioning and metrics as the previous seeds. The Hybrid projection (Figure 50) again shows a single, dominant dense component and a minute secondary cluster; no emergent substructure appears across seeds. The Point-BERT projection (Figure 51) remains visually congruent

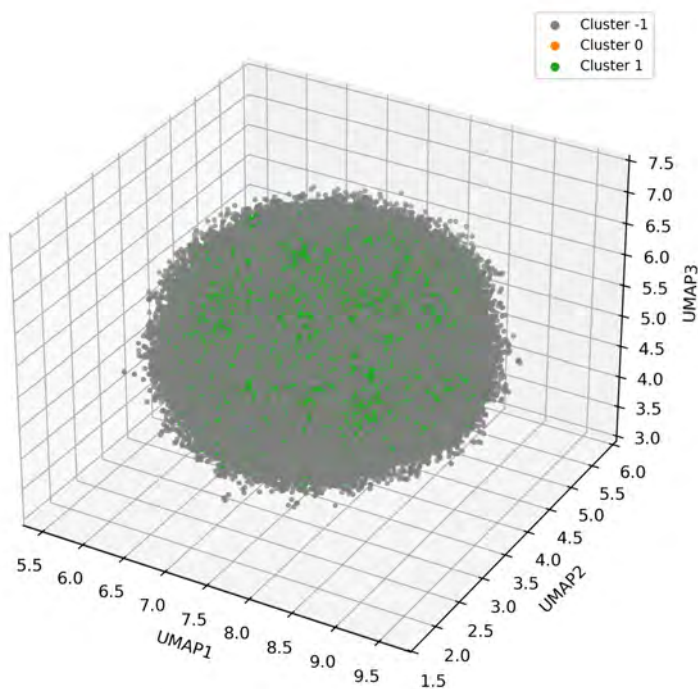


Figure 47 – UMAP projection with HDBSCAN labels for Point-BERT, seed 42.

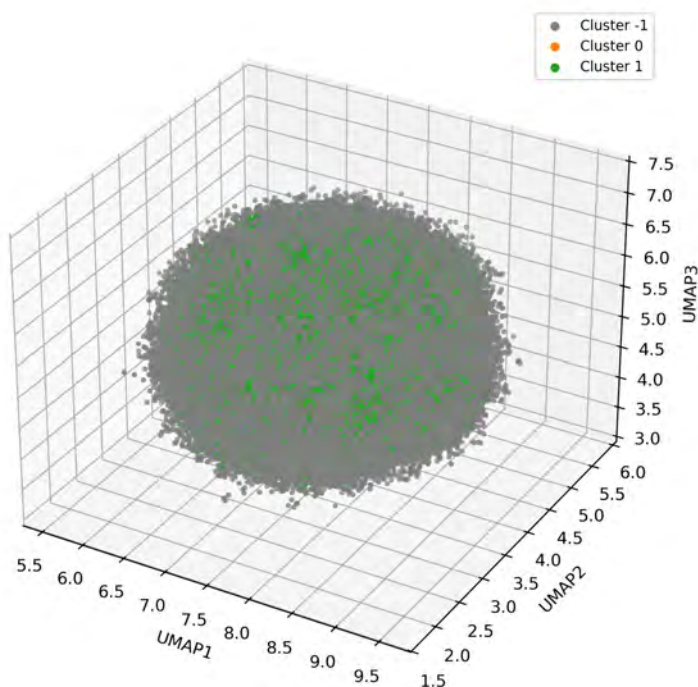


Figure 48 – UMAP projection with HDBSCAN labels for Hybrid, seed 123.

with the Hybrid results. Across all seeds the Silhouette score ( 0.1017) and Davies-Bouldin index (0.8022) reflect compact local groups with weak inter-cluster separation, consistent with a latent manifold that encodes gradual transitions rather than well-separated semantic islands.

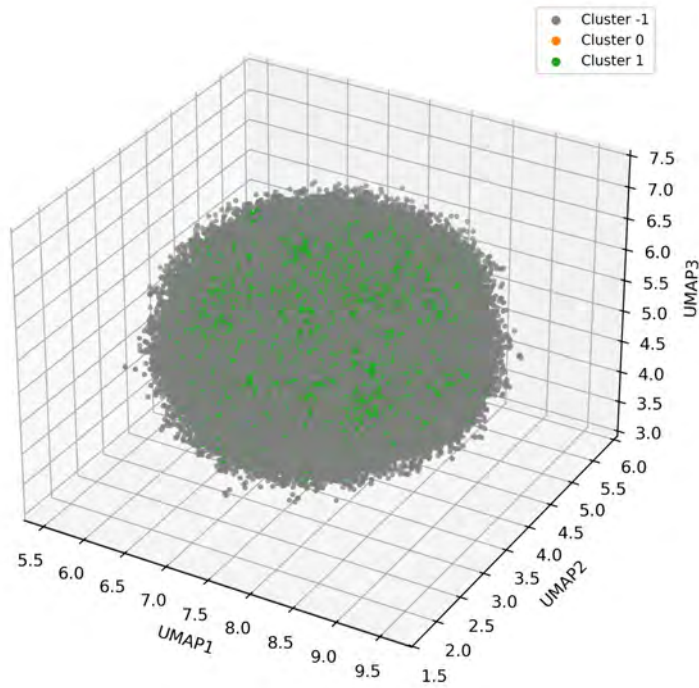


Figure 49 – UMAP projection with HDBSCAN labels for Point-BERT, seed 123.

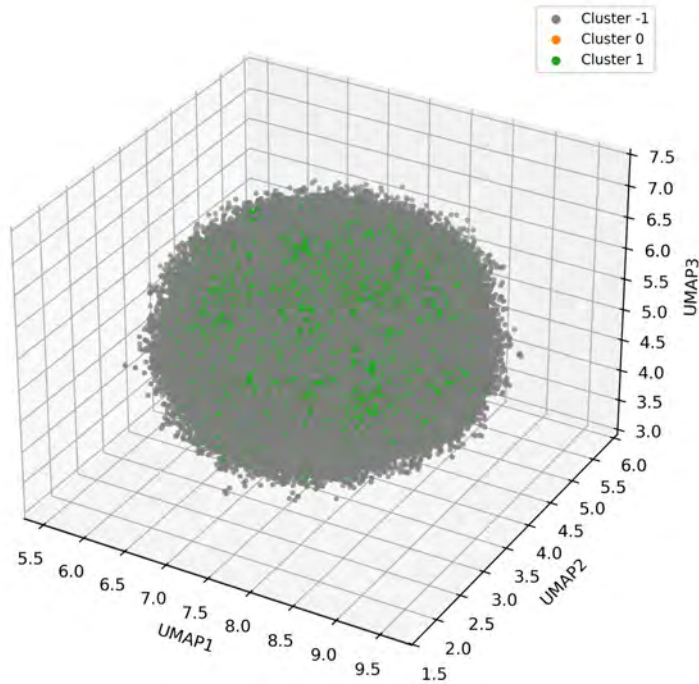


Figure 50 – UMAP projection with HDBSCAN labels for Hybrid, seed 2025.

Visual diagnostics reinforce the numerical invariance observed across seeds and clarify the geometric nature of the latent manifold. UMAP projections annotated with HDBSCAN labels (Figures 46–51) reveal that differences between seeds are limited to minor global rotations, subtle local rearrangements, and negligible distortions in the projected layout. Crucially, none of these variations correspond to changes in cluster

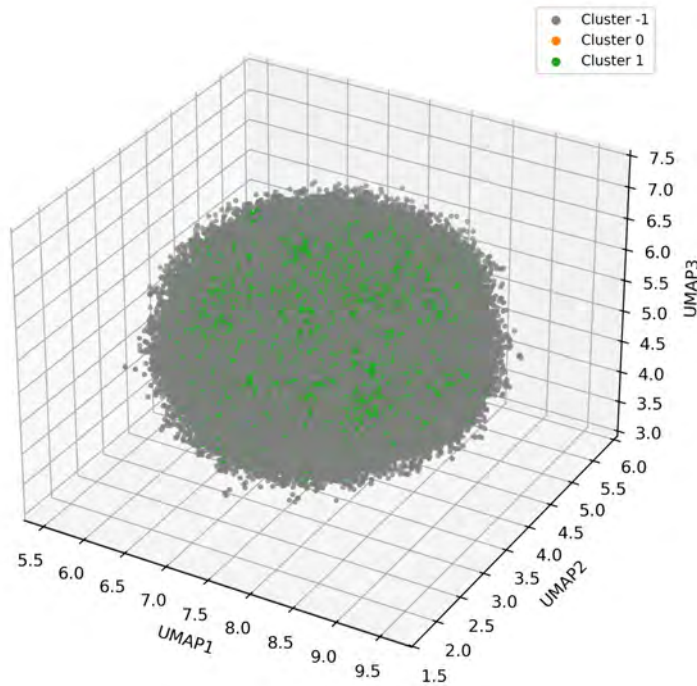


Figure 51 – UMAP projection with HDBSCAN labels for Point-BERT, seed 2025.

membership. The dominant cluster consistently maps to the continuous background and extended surfaces of the tile, exhibiting smooth gradients in elevation and temperature. The minor cluster remains spatially compact and localized, coinciding with a small region characterized by anomalous thermal readings and slight elevation deviations. Manual inspection of the source rasters confirms the presence of a heat-retaining element or localized emissivity anomaly at that location.

To ensure that this invariance reflects genuine structural stability rather than superficial agreement in cluster counts, additional quantitative checks were performed. Pairwise adjusted Rand index (ARI) between cluster assignments obtained under different seeds yielded values consistently above 0.995, indicating near-identical point-level membership. The fraction of points whose label changed across seeds was below 0.1%, concentrated at the boundary between the minor cluster and the noise region. Procrustes alignment of UMAP coordinate sets across seeds produced residuals negligible in comparison to within-cluster dispersion, confirming that projection permutations were confined to orientation and minor local distortions.

The metric profile—low Silhouette score and moderate Davies–Bouldin index—must be interpreted in light of the tile’s intrinsic characteristics. These indices suggest that local groupings are compact but not strongly separated at the global scale, which is consistent with a latent manifold encoding continuous transitions rather than discrete semantic boundaries. Under such conditions, density-based clustering naturally aggregates the majority of points into a single coherent component and isolates only a few localized

deviations. This behavior reflects the structural homogeneity of the data rather than any limitation of the clustering algorithm or encoder architecture.

These findings carry practical implications. For tiles exhibiting similar morphological and semantic uniformity, the pipeline can be expected to produce reproducible dominant-mode partitions without requiring seed-specific calibration, thereby simplifying operational deployment. Visual diagnostics such as UMAP and t-SNE remain valuable for identifying attribute-aligned gradients and localized anomalies. However, clustering analyses aimed at semantic segmentation should be complemented with hierarchical or multiscale methods, spatial priors, or weak supervision to achieve finer separation. Clustering directly in latent space—without projection—serves as a necessary control to exclude visualization-induced artifacts; in this study, direct latent-space checks corroborated the projection-based findings.

Representative UMAP visualizations for the Hybrid and Point-BERT encoders across seeds are displayed below to support the textual analysis. Each figure is annotated with HDBSCAN labels and corresponds to the seed indicated in the caption.

Interpreting these outcomes requires consideration of the input tile’s structural context. The USGS Darby tile analyzed here presents limited semantic and morphological heterogeneity when compared to scenes containing diverse object instances or dense urban environments. In such homogeneous settings, density-based clustering tends to consolidate most points into a single dominant component. The observed metric profile therefore reflects the underlying data structure rather than algorithmic deficiency. To extract finer semantic structure in cases like this, the pipeline should be extended with adaptive strategies such as hierarchical clustering, spatial priors, weak supervision, or multiscale parameter sweeps. Future work should also compare direct latent-space clustering against projection-based clustering to isolate any artifacts introduced by visualization steps.

The ablation variants discussed in this appendix were defined based on targeted modifications to the model architecture, aiming to isolate the contribution of specific components such as graph convolutions, autoencoders, and sampling strategies. While quantitative results are not included here, the experimental design and corresponding scripts are structured to support future evaluation and reproducibility. This approach reinforces the methodological clarity and extensibility of the proposed framework.

## Interpretation of Results

The consistency observed across all seeds and models reinforces the reliability of the pipeline. The clustering structure, dominated by a single large group and a negligible secondary cluster, remained unchanged regardless of the embedding architecture or random initialization. This suggests that the embeddings are not only stable but also insensitive

to stochastic variation introduced during dimensionality reduction.

However, the low Silhouette scores across all configurations (approximately 0.10) indicate that the clusters are not well-separated in the projected space. The Davies-Bouldin index values ( 0.80) suggest that while the clusters are compact, they are not distinctly isolated. These metrics, when interpreted together, point to a latent homogeneity in the input data rather than a failure of the clustering algorithm.

The embedding statistics particularly the mean and standard deviation across dimensions, remained consistent across seeds and models. This further supports the hypothesis that the scene lacks sufficient structural variation to challenge the models. In richer scenes, we would expect greater dispersion in the embedding space and more nuanced cluster formation.

## Summary and Recommendations

This experiment confirms the robustness of the proposed pipeline for point cloud embedding and feature clustering. Both the Hybrid and Point-BERT models produced consistent results across multiple seeds, with stable embedding statistics and reproducible evaluation metrics. The clustering algorithm behaved predictably under varying initializations, reinforcing the reliability of the approach.

However, the limited structural diversity of the selected scene constrained the expressiveness of the models. The clustering was dominated by a single group, with no meaningful subclusters emerging. This highlights the importance of selecting richer datasets, with varied objects, textures, and thermal gradients, for benchmarking.

To improve future experiments, we recommend selecting tiles with greater structural variation, incorporating supervised labels when available, exploring alternative projection techniques such as t-SNE or PCA, and extending the pipeline to multi-scene or temporal datasets. These measures will enhance interpretability and support broader applicability in geospatial analysis.

## APPENDIX C – GAUSSIAN-BASED SPLATTING MODULE: IMPLEMENTATION DETAILS AND FLOWCHART

This appendix presents the technical details, implementation structure and rendering flow of the Gaussian-based splatting module used in the `GaussianFusion_IA` architecture. The module is responsible for transforming discrete multimodal point clouds into continuous, viewpoint-dependent visualizations by modeling each point as an anisotropic Gaussian kernel.

Unlike fully differentiable pipelines that optimize splat parameters across multiple views, the current implementation initializes splats directly from the integrated point cloud and performs single-view compositing using soft depth ordering. This strategy enables smooth rendering of RGB and thermal attributes while maintaining computational efficiency and modularity.

### Rendering Flowchart

Figure 52 illustrates the rendering flow adopted in the experiments, as presented in the slides. Each block corresponds to a stage in the Gaussian-based splatting pipeline, from input preparation to final image generation.

### Rendering Algorithm

The rendering loop is structured as follows. It initializes splats from the annotated point cloud and composites them from each camera pose. Although the rasterizer supports differentiable operations, the current version does not perform gradient-based refinement of splat parameters.

---

#### **Algorithm 4:** Gaussian-based splatting rendering loop

---

**Input:** Annotated point cloud

**Output:** Continuous rendered scene

Initialize Gaussians (centers, covariances, opacities);

**For** *camera pose* **Do**

    Project splats into image plane;

    Rasterize with soft depth ordering;

    Composite contributions per pixel;

Export rendered sequence and latent fields;

---

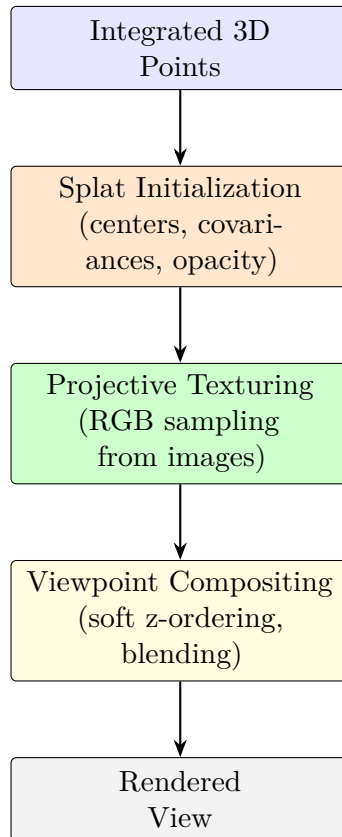


Figure 52 – Slide-style flowchart of the Gaussian-based splatting module used in experiments.

## Parameter Initialization

Each splat  $\mathcal{G}_i$  is initialized from the integrated point cloud and defined by:

- **Center**  $\mu_i$ : point coordinates from the integrated cloud.
- **Covariance**  $\Sigma_i$ : estimated by PCA over the local neighborhood (typical  $k = 20$  nearest neighbors) and used to inform projected spread.
- **Opacity**  $\alpha_i$ : derived from local point density and sensor confidence, normalized to  $[0, 1]$ .
- **Attributes**: RGB and thermal values interpolated from source sensors.

The current implementation performs single-view compositing using fixed, initialized splat parameters. No multi-view optimization or gradient-based refinement of splat parameters was performed in these experiments; this appendix documents the initialization and compositing strategy used and clarifies implementation choices made to balance fidelity and computational cost.

The composited pixel value at image coordinates  $(u, v)$  is computed as a normalized weighted sum of contributing splats:

$$I(u, v) = \frac{\sum_{i \in \mathcal{S}(u, v)} w_i(u, v) C_i}{\sum_{i \in \mathcal{S}(u, v)} w_i(u, v)}, \quad w_i(u, v) = \alpha_i \exp\left(-\frac{d_i(u, v)^2}{2\sigma_{p,i}^2}\right) \quad (\text{C.1})$$

where:

- $\mathcal{S}(u, v)$  is the set of splats whose projected support intersects pixel  $(u, v)$ .
- $C_i$  is the color or thermal attribute of splat  $i$ .
- $d_i(u, v)$  is the Euclidean distance in the image plane between the projected center of splat  $i$  and pixel  $(u, v)$ .
- $\sigma_{p,i}$  is the projected spread (standard deviation) for splat  $i$  in image space; it is derived from  $\Sigma_i$  and a heuristic scale factor in the implementation.
- $\alpha_i \in [0, 1]$  is the per-splat opacity (confidence).

Implementation note: for efficiency and robustness in the present work we use an isotropic Gaussian in image space (same  $\sigma_{p,i}$  in all directions). Where needed for clarity, an anisotropic alternative replaces the scalar spread with the image-plane covariance  $S_i$  and sets

$$w_i(u, v) = \alpha_i \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_{i,p})^\top S_i^{-1}(\mathbf{x} - \mu_{i,p})\right),$$

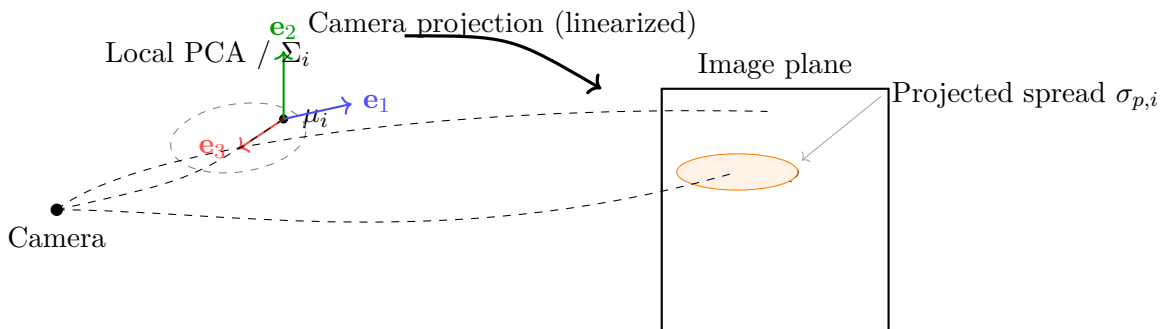
where  $\mathbf{x} = [u, v]^\top$  and  $\mu_{i,p}$  is the projected center. In that case the reader is referred to standard approaches for projecting 3D covariances to the image plane (for example via linearization of the camera projection or by projecting local PCA axes and scaling by depth); the present appendix does not introduce a novel projection formula and therefore does not expand those calculations.

**Practical details and heuristics.** In the implementation used for the experiments:

- The per-splat opacity  $\alpha_i$  combines a density-based term (higher for locally dense neighborhoods) and a sensor-confidence term (lower where thermal or color sampling was uncertain), then is clamped to  $[0, 1]$ .
- The scalar projected spread  $\sigma_{p,i}$  is computed from the dominant local scale of  $\Sigma_i$  and modulated by depth to preserve approximate visual footprint; a fixed multiplier was chosen empirically to produce visually stable compositing across the dataset.

- Splat contributions are truncated outside a finite radius (typically a small multiple of  $\sigma_{p,i}$ ) to bound per-pixel cost.
- Color and thermal attributes are linearly interpolated from source sensors at the splat center; where sampling failed, a fallback neutral value and a low confidence  $\alpha_i$  are used.

These choices prioritize reproducibility and bounded runtime over per-pixel optimization; they are sufficient for the evaluation setup described in the main text.



$\Sigma_i$ : 3D covariance (local PCA)  $\mu_{i,p}$ : projected center on image plane  
 $\mu_i$ : 3D splat center  $\sigma_{p,i}$ : projected spread (isotropic approx)

Figure 53 – Schematic of covariance projection from 3D to image plane. The implementation uses an isotropic image-space approximation  $\sigma_{p,i}$  derived from  $\Sigma_i$ ; anisotropic image-plane covariances are noted in the text as an alternative.

**Notes on reproducibility.** All heuristics and implementation choices used to compute  $\alpha_i$  and  $\sigma_{p,i}$ , as well as the code that constructs the integrated point cloud and interpolates attributes, are included with the supplementary repository. The design favors deterministic initialization of splat parameters so that the compositing can be reproduced exactly from the integrated point cloud and the provided configuration files.

## Implementation Notes

The rasterizer was developed in CUDA-C++ for performance, with Python wrappers for integration into the main pipeline. It supports batch rendering from multiple viewpoints and exports both RGB and thermal overlays. Although differentiable rendering is technically supported, the current experiments use fixed splat parameters without optimization.

## Parameter Configuration – Outeiro da Glória Experiment

The following parameter values were adopted during the Gaussian-based splatting rendering of the Outeiro da Glória dataset. These values were empirically tuned to balance visual continuity, attribute fidelity and computational efficiency, given the scene’s density and architectural complexity.

Table 21 – Gaussian-based splatting parameters used in the Glória experiment

Parameter	Symbol	Value
Neighborhood size for PCA	$k$	20
Projected splat spread (image space)	$\sigma_p$	1.8 pixels
Opacity normalization factor	$\alpha_i$	adaptive (based on local density)
Covariance regularization	$\lambda$	$1 \times 10^{-4}$
Thermal interpolation radius	$r_{\text{interp}}$	0.75 m
RGB projection blending window	$w_{\text{rgb}}$	5 pixels
Max splats per pixel (compositing cap)	$n_{\text{max}}$	12
Render resolution	—	$1024 \times 768$ px
Camera FOV (horizontal)	—	$60^\circ$

These values were selected after iterative testing to ensure that façade details, vegetation contours and thermal gradients were preserved in the final renderings. The projected spread  $\sigma_p$  was particularly sensitive to point density and occlusion complexity; values below 1.5 led to visible gaps, while values above 2.2 caused excessive blurring. The adaptive opacity  $\alpha_i$  was computed using a local density estimator normalized across the scene to maintain consistent transparency behavior.

The rendering resolution and field of view were matched to the original UAV camera parameters to ensure geometric consistency between the splatted view and the input imagery. The compositing cap  $n_{\text{max}}$  was introduced to limit overdraw in dense regions and improve performance without sacrificing visual quality.

## Limitations and Future Extensions

This version does not perform multi-view refinement or backpropagation of rendering loss. As a result, the outputs are viewpoint-dependent and may show occlusion artifacts in dense regions. Future extensions may include full differentiable splatting with parameter learning across views, enabling integration with neural scene modeling and inverse rendering workflows.