A Two-Stage Diffusion Framework for 3D Building Reconstruction from Single Nadir Images

Youssef Korny

A Thesis Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements
for the Degree of Master of science

Graduate Program in Earth and Space Science

York University
Toronto, Ontario

June, 2026
©Youssef Korny, 2026

Abstract

Reconstructing 3D urban geometry from single nadir aerial images is an ill-posed problem due to severe vertical ambiguity and occluded building facades. To resolve this, this thesis proposes GeoPriorPC, a novel two-stage generative framework bridging foundational 2D image priors and unstructured 3D point clouds. Stage 1 fine-tunes a Latent Diffusion Model (LDM) to extract a compressed geometric prior, specifically, oblique surface normals, from flat nadir inputs. Stage 2 employs a Diffusion Point Transformer (DiPT) backbone. It utilizes a dual-conditioning mechanism, integrating global adaptive Layer Normalization (adaLN-Zero) and local spatial cross-attention, to structurally guide the 3D denoising process using the learned 2D prior. Evaluated on the AIM2PC benchmark, GeoPriorPC establishes a new state-of-the-art, achieving a 44% relative improvement in F1-score and reducing Earth Mover's Distance by 49%. By explicitly resolving depth ambiguity, this framework successfully synthesizes the accurate architectural manifolds critical for urban digital twins.

Contents

Page

Abstract **ii**

Contents **iii**

List of Tables **vi**

List of Figures **vii**

List of Abbreviations **ix**

1 Introduction **1**

1.1 Motivation 1

1.2 The Representation Gap: From Point Clouds to Parameterized Models 2

1.3 The Problem Statement: The Ambiguity of the Nadir View 2

1.4 The Proposed Solution: GeoPriorPC 3

1.4.1 Stage 1: Latent Geometric Prior Learning 3

1.4.2 Stage 2: Dual-Conditioned 3D Generation 4

1.5 Research Objectives and Contributions 5

1.6 Thesis Organization 5

2 Literature Review **7**

2.1 Single-View 3D Reconstruction in Remote Sensing 7

2.1.1 The 2.5D Regression Era 7

2.1.2 Structured Representations and the Topological Bottleneck 7

2.1.3 The Generative Point Cloud Era 11

2.2 Deep Learning Architectures for 3D Point Clouds 11

2.2.1 The Point-Based models: MLPs and Spatial Querying 12

2.2.2 The Voxel-Based models: Sparse Convolutions 12

2.2.3 Transformers: Serialization and Global Attention 13

2.3 Data Modalities and Urban Benchmarks 14

2.3.1 The Reliance on Active 3D Sensors 14

2.3.2 Bridging the Modality Gap with Generative Benchmarks 14

2.4 The Evolution of Generative Models 15

2.4.1 Pre-Diffusion Generative Baselines 15

2.4.2 The Mathematical Foundation of Diffusion 15

2.4.3 Latent Compression and 3D Adaptation 16

2.5 Generative Priors for 3D Geometry 17

2.5.1 The Implicit 3D Knowledge of Foundation Models 17

2.5.2 Extracting the Prior: Optimization vs. Feed-Forward . . . 18
2.6 The Gap in the Literature . . . 18

3 GeoPriorPC: A two-stage diffusion model for 3D building reconstruction 20
3.1 Overview of the Architecture . . . 20
3.2 Mathematical Foundations of Diffusion Models . . . 20
3.2.1 2D Latent Diffusion Models . . . 21
3.2.2 Point Cloud Diffusion: Operating in the Direct Coordinate Space . . . 22
3.3 Problem Formulation and Input Representation . . . 22
3.3.1 Input Space Definition . . . 22
3.3.2 Target Output Space . . . 23
3.3.3 The Formulation of the Geometric Prior . . . 23
3.4 Stage 1: Geometric Prior Extraction and Data Engineering . . . 23
3.4.1 Data Preprocessing and Oblique Normal Rendering . . . 23
3.4.2 Latent Space Encoding and Computational Optimization . . . 24
3.4.3 Latent Diffusion Fine-Tuning Dynamics . . . 25
3.4.4 Architectural Explorations . . . 26
3.5 Stage 2: Geometry-Conditioned 3D Point Generation . . . 29
3.5.1 Multi-Modal 2D Feature Extraction and Projection . . . 29
3.5.2 The Diffusion Point Transformer (DiPT) Backbone . . . 29
3.5.3 Dual-Conditioning Mechanism for Geometric Priors . . . 31
3.5.4 Ablation Model Variations . . . 32

4 Experiments and Results 34
4.1 Experimental Setup . . . 34
4.1.1 Dataset and Preprocessing . . . 34
4.1.2 Implementation and Training Details . . . 34
4.1.3 Evaluation Metrics . . . 35
4.2 Main Quantitative Results . . . 36
4.2.1 Global Metric Comparison . . . 36
4.2.2 Performance Analysis . . . 36
4.3 Qualitative Evaluation . . . 37
4.3.1 Visual Comparisons and Structural Fidelity . . . 37
4.4 Complexity and Inference Efficiency . . . 37
4.5 Ablation Studies and Architectural Explorations . . . 38
4.5.1 Impact of Pre-trained Latent Priors (Stage 1) . . . 38
4.5.2 Backbone and Conditioning Strategies (Stage 2) . . . 40
4.5.3 Architectural Exploration: Single-View vs. Multi-View Conditioning . . . 40
4.5.4 Architectural Exploration: Isotropic vs. Hierarchical DiPT Backbone . . . 42
4.5.5 Architectural Exploration: Explicit Aspect Ratio Conditioning . . . 43

4.6 Architectural Exploration: Effect of Random Seeds in the Multi-View Setting 45
4.6.1 Stochasticity in Stage 2 (Image-to-Point Generation) 45
4.6.2 Stochasticity in Stage 1 (Latent Geometric Prior Extraction) 45
4.6.3 Addressing the Upper Bound: Test-Time Ensembling 46

5 Limitations **47**
5.1 Cross-View Geometric Inconsistency 48
5.2 Generative Stochasticity and the "Oracle" Gap 48
5.3 Dataset Bias and Inference Overhead 48
5.4 The Transitional Nature of Unstructured Representations 48

6 Conclusion and Future Work **50**
6.1 Conclusion 50
6.2 Future Work 50
6.2.1 Scaling Data Diversity and Mitigating Propagation Errors 51
6.2.2 Harnessing the Generative Oracle: Uncertainty-Aware Routing 51
6.2.3 Enforcing Epipolar Consistency in Multi-View Generation 51
6.2.4 Stabilizing Hierarchical Transformers for Diffusion Dynamics 52
6.2.5 Inference Acceleration via Diffusion Distillation 52
6.2.6 Bridging the Modality Gap to Parametric B-rep Modeling 53

Bibliography **54**

List of Tables

1 Chronological evolution of deep learning architectures for 3D point clouds, highlighting their mechanisms and limitations for large-scale geospatial reconstruction. 12
2 Hyperparameters for Data Rendering and Stage 1 LDM Fine-Tuning. 27
3 Quantitative comparison of single-image 3D building reconstruction on the AIM2PC dataset [1]. Arrows indicate whether higher (↑) or lower (↓) scores are optimal. Best results are highlighted in bold. 36
4 Comparison of model complexity and average iteration time evaluated on an NVIDIA L40S GPU. 38
5 Evaluation of Stage 1 normal map prediction quality, comparing random initialization versus fine-tuning pre-trained LDM weights. 40
6 Ablation study on architectural backbones and prior conditioning strategies. 41
7 Quantitative comparison of Single-View and Multi-View conditioning. The ”Oracle” column represents the theoretical maximum if the best prediction per sample is dynamically selected, proving the complementary nature of the two methods. 41
8 Impact of explicit aspect ratio embedding on Stage 1 and Stage 2 performance. 45
9 Effect of Random Seeds on Stage 2 Inference 45
10 Effect of Random Seeds on Stage 1 Prior Generation 46
11 Comparison of Ensembling vs. Oracle Upper Bound 46

List of Figures

1 **Overview of the two-stage image-to-point diffusion framework.** The left panel illustrates our pipeline: a baseline trained from scratch (top) is compared to our approach, which uses a fine-tuned Latent Diffusion Model (LDM) to extract geometric priors (Stage 1) to guide 3D point cloud generation (Stage 2). The right panel shows qualitative comparisons of our results (Ours) against the ground truth (GT). 4

2 An illustration to show the difference between wireframe on the left, mesh in the middle, and point clouds on the right 9

3 Samples of the generated dataset for LDM training. Each sample contains an RGB image with mask subtraction and a target normal oblique image. All processing and rendering run on the fly during training. 24

4 The forward training process of the geometric prior LDM. Gaussian noise is added to the ground-truth normal latent, concatenated with the encoded RGB image latent, and passed through the U-Net to compute the noise prediction loss. 26

5 The accelerated inference process using Denoising Diffusion Implicit Model (DDIM) sampling. The model denoises pure Gaussian noise over 50 steps, heavily conditioned on the nadir image latent, to output the final compressed geometric prior $\hat{z}_0$. 27

6 Explanation of aspect ratio injection. We are using an Multi-layer Perceptron (MLP) to create an embedding with the same size as the time embedding, which is then added together and fed to the model via cross-attention. 28

7 An input and output example from the multiview LDM. The model is now expecting to receive the same image duplicated 4 times, and for each duplication, the model is predicting a different oblique normal view. 29

8 The proposed GeoPriorPC main. On the left, the model architecture is based on Diffusion Point Transformer (DiPT), while on the right, an illustration of the serialized process used in serialized attention. 30

9 A diagram explaining the structural differences between the Isotropic and U-Net designs. a) Represents the isotropic design, where we utilize N sequential PTv3 layers. Both the feature dimension (dim) and the total number of point tokens (N_p) remain constant throughout the network. b) Represents the U-Net design, where the number of tokens (N_p) and the feature dimension dynamically change across layers through serialized spatial pooling and unpooling operations. 31

10 Comparison of the ablation model designs: (a) GeoPriorPC-anaLN handles the latent normal map prior solely as a compressed global context vector via adaLN-Zero layers without local cross-attention, while (b) GeoPriorPC-Ignore completely bypasses the geometric prior stage and relies strictly on the raw 2D input maps. 33

11 Qualitative results of finetuned LDM. It shows that the model can successfully generate the normal oblique view from just a single nadir image. 38

12 Qualitative comparison of 3D building reconstructions. From left to right: Input nadir image, Ground Truth, Our model with global context only, our full GeoPriorPC model, and AIM2PC baseline. 39
13 Results from an LDM trained from scratch. The results show that the model completely collapsed and couldn't generate any meaningful output. 40
14 A scatter plot showing F1-score for both single and multiview predictions. Each point represents a test sample, while the x-axis represents the f1-score for the single-view model and the y-axis represents the f1-score for the multiview model. 42
15 A scatter plot showing F1-score for both single and multiview predictions. Each point represents a test sample, while the x-axis represents the f1-score for the single-view model and the y-axis represents the f1-score for the multiview model. 43
16 A comparison of training dynamics between the hierarchical U-Net and the Isotropic designs. (a) Displays the clipped gradient norm across training steps, highlighting the severe gradient explosion and volatility inherent to the U-Net design compared to the bounded stability of the isotropic model. (b) Illustrates the resulting impact on model convergence. While the U-Net shows a strong initial decrease in error, the subsequent gradient instability prevents it from effectively refining the geometric coordinates, whereas the isotropic architecture smoothly minimizes the Earth Mover's Distance. 44
17 Failure cases demonstrating error propagation. Geometric inaccuracies in the LDM-predicted normal maps dictate the structural artifacts in the final 3D point clouds. 47

List of Abbreviations

A/H	Area-to-Height Ratio
B-rep	Boundary Representation
BIM	Building Information Modeling
CAD	Computer-Aided Design
CFD	Computational Fluid Dynamic
CVPR	Conference on Computer Vision and Pattern Recognition
DiPT	Diffusion Point Transformer
IoT	Internet of Thing
LDM	Latent Diffusion Model
LoD1	Level-of-Detail 1
PTv3	Point Transformer v3
PTv1	Point Transformer v1
PVCNN	Point-Voxel Convolutional Neural Network
UDT	Urban Digital Twin
adaLN-Zero	adaptive Layer Normalization
CNN	Convolutional Neural Network
DDPM	Denoising Diffusion Probabilistic Model
SOTA	state-of-the-art
MLP	Multi-layer Perceptron
NURBS	Non-Uniform Rational B-Splines
FPS	Farthest Point Sampling
KNN	K-Nearest Neighbor
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
AE	Autoencoder

ELBO	Evidence Lower Bound
MSE	Mean Squared Error
DDIM	Denoising Diffusion Implicit Model
SDS	Score Distillation Sampling
cGAN	Conditional Generative Adversarial Network
CDPM	Centered Denoising Probabilistic Model
ViT	Vision Transformer
EMD	Earth Mover's Distance
DCD	Density-Aware Chamfer Distance
MAE	Mean Absolute Error
DSM	Digital Surface Model
GIS	Geographic Information System

1 Introduction

1.1 Motivation

The rapid urbanization of the global population has transformed modern cities into highly complex, interconnected ecosystems. To effectively manage these evolving environments and transition towards "Smarter Cities", urban planners and policymakers increasingly rely on 3D city models and the emerging paradigm of Urban Digital Twins (UDTs) [2]. Unlike traditional 2D maps, UDTs provide dynamic, virtual replicas of the physical built environment, serving as the foundational infrastructure for a smarter city. They enable advanced spatial analysis, real-time Internet of Things (IoTs) monitoring, autonomous vehicle navigation, and data-driven simulation across the entire urban lifecycle [3].

The demand for comprehensive 3D urban geometry is driven by its critical role across several high-impact domains where 2D representations are fundamentally inadequate:

- **Smarter Environmental and Microclimate Simulations:** Accurate 3D building morphology, including precise height, volume, and roof topologies, is essential for modeling urban microclimates. Vertical structural parameters directly govern Computational Fluid Dynamics (CFDs) simulations for assessing pedestrian wind comfort, calculating the attenuation of urban noise propagation, and estimating building-level energy demands and solar potential [4].
- **Urban Planning and Spatial Analysis:** 3D geometry enables the calculation of complex morphological indicators, such as the Area-to-Height Ratios (A/Hs) and vertical compactness. These metrics are vital for assessing urban density, skyline continuity, and land-use optimization, ensuring that new developments integrate seamlessly with existing urban fabrics [5, 6].
- **Disaster Management and Resilience:** In the context of climate change and extreme weather events, true 3D manifolds are required to simulate disaster scenarios accurately. High-resolution building geometries are critical for visualizing the impact of hurricane storm surges on urban infrastructure [7] and developing highly efficient, staged flood evacuation systems within complex multi-exit environments.

Despite the necessity for these geometric models, a significant bottleneck exists in the data acquisition process. Traditionally, high-fidelity 3D building models are generated using airborne LiDAR or multi-view photogrammetry [8]. While these capture precise coordinates, they are prohibitively expensive and logistically complex for global-scale, high-frequency updates. Consequently, the world's 3D urban data quickly becomes outdated as cities rapidly evolve.

In contrast, nadir (top-down) satellite and aerial imagery is abundant, globally accessible, and frequently updated. The core motivation of this thesis is to bridge the gap between the abundance of 2D aerial imagery and the critical demand for 3D urban geometry, developing a generative methodology capable of reliably inferring the missing vertical dimension from a single top-down perspective.

1.2 The Representation Gap: From Point Clouds to Parameterized Models

When synthesizing 3D structures from 2D images, the choice of the output data representation dictates both the flexibility of the generative model and its utility for downstream smart city systems. Many large-scale reconstruction efforts focus on generating Level-of-Detail 1 (LoD1) models by extruding building footprints based on estimated heights [9]. While computationally efficient, these simple prismatic representations fail to capture complex roof structures, dormers, and multi-level facades.

To overcome these strict topological constraints, this thesis adopts the 3D point cloud as the primary generative medium. Point clouds offer a highly flexible, unstructured format that can represent arbitrarily complex, non-manifold geometries without the restrictive grid limitations of standard surface models [1].

However, raw point clouds carry inherent drawbacks for smart city integration. A point cloud is essentially a "smallest common denominator" format; it is merely an unorganized array of spatial coordinates that lacks topological connectivity, semantic constraints, and explicit surface boundaries [10]. Consequently, raw point clouds are highly redundant and notoriously difficult for downstream urban simulation systems and Computer-Aided Design (CAD) software to interpret natively without extensive reverse engineering [11].

Ultimately, the ideal geometric representations for smart city systems and Building Information Modeling (BIM) are highly structured, parametric models such as wireframes, meshes, or Boundary Representation (B-rep) [6]. These formats define buildings explicitly through mathematical planes, vertices, and watertight edges, making them easy to manipulate, edit, and run in physical simulations [12].

Directly generating complete B-rep or parametric CAD models from single nadir images remains an extraordinarily difficult task. This difficulty stems from the fundamental mismatch between the discrete, non-linear topological constraints inherent to CAD representations and the continuous, differentiable optimization paradigms utilized by neural networks. Therefore, generating highly accurate, dense point clouds serves as the critical, necessary stepping stone [13]. By successfully resolving the vertical ambiguity of the nadir view and outputting a robust 3D point cloud, the framework proposed in this thesis provides the foundational geometry required. It enables subsequent pipelines to apply plane-fitting algorithms, structural parsing, and reverse-engineering techniques, thereby bridging the gap between raw 2D pixels and the parameterized 3D models required by smarter cities [11, 10].

1.3 The Problem Statement: The Ambiguity of the Nadir View

The fundamental challenge addressed in this thesis is the inherently ill-posed nature of single-view 3D reconstruction, specifically when operating from a nadir (top-down) perspective. To construct a 3D point cloud of a building from a 2D aerial image, an algorithm must solve an inverse projection problem: mapping a 2D pixel coordinate (u, v) back into its corresponding 3D spatial coordinate (x, y, z).

In standard multi-view photogrammetry, this inverse problem is solved deterministically using epipolar geometry, where the parallax between multiple overlapping images allows for precise depth triangulation. However, when constrained to a single perspective, the depth dimension (z) is entirely lost during the forward projection process.

This loss of dimension creates a mathematically "many-to-one" mapping problem. A single 2D top-down image can theoretically be produced by an infinite number of different 3D shapes. For example, from a

strict nadir perspective lacking distinct lighting or shadows, a building with a completely flat roof may yield an identical 2D footprint and pixel array as a building with a highly complex, multi-tiered pitched roof.

Furthermore, the nadir perspective suffers from severe geometric occlusion. Because the camera is positioned directly above the target, the vertical planes of the building, namely, the facades, windows, doors, and structural undercuts, are entirely hidden from view. Without observable data for these vertical elements, deterministic regression models cannot accurately reconstruct the building's true manifold. When forced to predict a specific output, traditional deterministic models (such as standard convolutional architectures predicting depth maps) tend to average out the uncertainty. This results in smoothed, low-fidelity geometries that fail to capture sharp roof ridges or precise 90-degree facade angles.

While human observers can often intuitively infer the 3D structure of a building from a single aerial image by interpreting contextual clues, such as shadows, architectural styles, and neighborhood layouts, computer vision models lack this intrinsic structural knowledge. To correctly infer the unobserved z-axis and the occluded vertical facades, a model must possess a strong structural prior: a learned understanding of how buildings are geometrically constructed in the real world.

Therefore, the problem is not merely a task of pixel translation, but of conditional generation under extreme uncertainty. Resolving the nadir ambiguity requires a framework capable of synthesizing missing topological information by sampling from a distribution of plausible 3D building shapes, conditioned on the partial evidence provided by the 2D image.

1.4 The Proposed Solution: GeoPriorPC

To overcome the inherent geometric ambiguities of the nadir perspective, this thesis introduces GeoPriorPC, a novel two-stage generative framework designed to reconstruct accurate 3D building point clouds from single aerial images. Rather than forcing a single neural network to blindly guess the missing out-of-plane geometry, GeoPriorPC systematically decomposes the reconstruction task into two distinct generative processes: (1) explicit geometric prior extraction and (2) geometry-conditioned 3D point cloud synthesis.

1.4.1 Stage 1: Latent Geometric Prior Learning

Because foundational text-to-image Latent Diffusion Models (LDMs), such as Stable Diffusion [14], have been trained on billions of internet images, their latent spaces inherently encode a generalized linear understanding of physical 3D properties [15]. The first stage of GeoPriorPC adapts this vast, pre-trained world knowledge for geospatial applications.

By fine-tuning an LDM via the Marigold framework [16], the model learns to translate a flat, top-down RGB image into a highly structured oblique surface normal map. This intermediate map explicitly models the varying angles of the roof faces and vertical facades. Crucially, rather than decoding this normal map back into high-resolution pixel space, which incurs heavy computational overhead, GeoPriorPC retains the output in its compressed latent space. This latent representation encodes global structural topology and relative surface orientations; by constraining the permissible 3D geometries to those consistent with these predicted surface normals, the framework effectively mitigates the inherent monocular depth ambiguity described in Section 1.3 without losing high-frequency structural details.

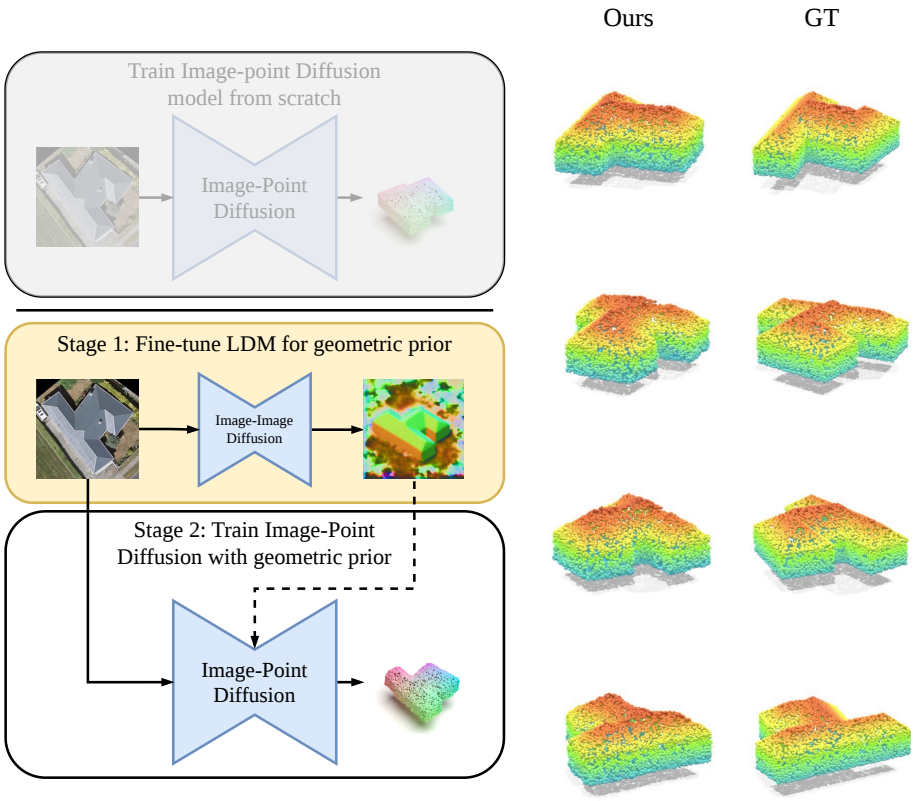


Figure 1: **Overview of the two-stage image-to-point diffusion framework.** The left panel illustrates our pipeline: a baseline trained from scratch (top) is compared to our approach, which uses a fine-tuned LDM to extract geometric priors (Stage 1) to guide 3D point cloud generation (Stage 2). The right panel shows qualitative comparisons of our results (Ours) against the ground truth (GT).

1.4.2 Stage 2: Dual-Conditioned 3D Generation

With the structural ambiguity resolved by the latent prior, the second stage focuses on generating the flexible, unstructured 3D representation. For this, GeoPriorPC employs a DiPT [17] backbone, heavily inspired by Point Transformer v3 (PTv3) [18]. Unlike traditional sparse 3D convolutional networks that suffer from localized receptive fields, the PTv3 backbone serializes the 3D points, allowing for highly efficient, patch-based global attention across the entire building structure.

To effectively map the 2D geometric prior into the 3D generation process, GeoPriorPC introduces a dual-conditioning mechanism:

1. **Global Modulation:** The latent normal map is compressed into a global context vector and injected into the diffusion blocks via adaptive Layer Normalization (adaLN-Zero). This ensures that the overall structural proportions and building scale are maintained while keeping early training convergence highly stable.

2. **Local Cross-Attention:** To capture intricate architectural details like specific roof ridges and sharp facade corners, the network flattens the 2D latent normal prior into sequence tokens. A spatial cross-

attention mechanism allows the denoising 3D point queries to ”attend” directly to these localized 2D geometric tokens, forcing the point cloud to align physically with the predicted surface angles.

By isolating the extraction of the structural prior from the generation of the 3D coordinates, GeoPriorPC effectively bridges the representation gap. It leverages the raw generative power of foundational image models while strictly guiding the 3D topology, yielding point clouds that respect accurate architectural manifolds.

1.5 Research Objectives and Contributions

The primary objective of this thesis is to develop a robust, generative framework capable of synthesizing highly accurate 3D building point clouds from single nadir-view aerial images. By addressing the fundamental geometric ambiguities of the top-down perspective, this research aims to provide a scalable pathway for extracting complex 3D urban structures from abundant 2D satellite and aerial imagery.

To achieve this objective, this thesis introduces GeoPriorPC, a novel two-stage diffusion pipeline. The specific technical and methodological contributions of this work are summarized as follows:

- **Foundational Model Adaptation for Geospatial Priors:** This thesis presents a novel adaptation of pre-trained LDMs for single-view 3D urban reconstruction. By fine-tuning the model to predict explicit oblique surface normals from a nadir input, the framework successfully extracts high-level structural knowledge from foundational image models without the computational overhead of pixel-space decoding.
- **Architectural Advancement in 3D Point Cloud Generation:** This research demonstrates that sequence-based, Transformer-driven diffusion backbones—specifically the DiPT utilizing a PTv3 architecture offer significant performance gains over traditional, voxelized CNN-based approaches (such as Point-Voxel Convolutional Neural Network (PVCNN)) for complex geospatial reconstruction tasks.
- **Robust Multi-Modal Conditioning Strategy:** A novel dual-conditioning mechanism is proposed to bridge the 2D prior and the 3D spatial generation. By integrating adaLN-Zero with a spatially-aware local cross-attention module, the 3D denoising process successfully attends to specific structural cues within the generated 2D normal maps. This strategy achieves state-of-the-art structural accuracy on existing urban datasets, reducing geometric anomalies and preserving uniform building topology.

The core methodology, the architectural design of GeoPriorPC, and most of the empirical findings presented in this thesis have been peer-reviewed and accepted for publication at the **3rd Urban Scene Modeling Workshop at the 2026 IEEE/CVF Conference on Conference on Computer Vision and Pattern Recognition (CVPR).**

1.6 Thesis Organization

The remainder of this thesis is organized as follows:

1. **Chapter 2:** Literature Review traces the evolution of single-view 3D reconstruction and establishes the mathematical foundation of diffusion models, identifying the critical gap in utilizing 2D foundation priors for 3D generation.

2. **Chapter 3:** Proposed Methodology details the theoretical formulation and architectural design of the two-stage GeoPriorPC framework, covering latent geometric prior extraction and the dual-conditioned 3D point cloud generation.

3. **Chapter 4:** Experiments and Results presents the empirical setup, dataset preparation, and a comprehensive quantitative and qualitative evaluation of GeoPriorPC against state-of-the-art baselines, supported by architectural ablation studies.

4. **Chapter 5:** Discussion and Limitations provides a critical analysis of the experimental outcomes, specifically addressing challenges regarding model architecture stability and the constraints of the training data distribution.

5. **Chapter 6:** Conclusion and Future Work summarizes the primary findings and technical contributions and proposes avenues for future research in smart city and Digital Twin applications.

2 Literature Review

2.1 Single-View 3D Reconstruction in Remote Sensing

The fundamental problem of extracting three-dimensional geometric structures from two-dimensional imagery is an inherently ill-posed inverse problem. When a 3D scene is projected onto a 2D sensor plane, the depth dimension is mathematically collapsed, resulting in an infinite number of plausible topographic configurations that could correspond to a single monocular observation. Historically, this geometric ambiguity was resolved by enforcing multi-view stereopsis or through active sensors like LiDAR, which are heavily constrained by prohibitive acquisition costs and low temporal revisit frequencies. Consequently, the geospatial vision community has relentlessly pursued single-view 3D reconstruction. This pursuit has evolved through a continuous effort to navigate the "trilemma" of geometric accuracy, topological validity, and computational efficiency, transitioning through three distinct representational eras.

2.1.1 The 2.5D Regression Era

Prior to the deep learning renaissance, monocular depth estimation relied exclusively on deterministic geometric assumptions and probabilistic graphical models, pioneered by frameworks like Make3D [19], which used Markov Random Fields to model coplanarity in superpixels. The paradigm rapidly shifted with the advent of multi-scale Convolutional Neural Networks (CNNs). Eigen et al. [20] proved that deep representation learning could automatically extract relative depth maps from single images, while subsequent work [21] established that simultaneously predicting depth alongside surface normals significantly improved structural boundary sharpness. As network architectures deepened, the introduction of residual skip connections and up-projection modules [22] became mandatory for preserving fine-grained spatial details during the decoding phase.

The remote sensing community quickly adapted these encoder-decoder architectures to predict absolute metric heights rather than relative depth. Srivastava et al. [23] introduced one of the first multi-task networks that jointly performed semantic segmentation and metric height estimation, proving that semantic priors are essential for resolving vertical ambiguities in orthographic imagery. This was fully standardized by Mou and Zhu [24], who successfully mapped aerial imagery to continuous 2.5D Digital Surface Models (DSMs) using a fully residual convolutional-deconvolutional network. However, networks trained with standard regression losses inherently produced over-smoothed, blob-like elevation predictions when faced with uncertainty. To combat this, researchers like Ghamisi and Yokoya [25] began utilizing Conditional Generative Adversarial Networks (cGANs) to hallucinate high-frequency topographic distributions. Despite these improvements, 2.5D rasters remain mathematically constrained; by assigning a single elevation value to each (x, y) coordinate, it is impossible to represent co-planar vertical points, meaning continuous regression cannot differentiate between a sloped hill and a sheer vertical facade.

2.1.2 Structured Representations and the Topological Bottleneck

To overcome rasterization artifacts and align with Geographic Information System (GIS) standards, the field shifted toward generating structured 3D representations. However, predicting topologically strict models

directly from single-view 2D observations, particularly from nadir and aerial perspectives, is mathematically ill-posed due to severe information loss during the 3D-to-2D projection process. Current research demonstrates that generating these highly parameterized models requires navigating a complex "trilemma" of geometric accuracy, topological validity, and computational efficiency.

The pursuit of structured urban reconstruction has primarily explored four representational paradigms, each encountering distinct topological bottlenecks when operating without an intermediate 3D geometric scaffold:

Procedural Extrusion and LoD1 Models LoD1 models represent the most basic architectural abstraction, extruding 2D building footprints to a uniform scalar height [26]. Recent advancements have revolutionized this approach through off-nadir image analysis. For instance, the SingleRecon[9] pipeline utilizes the ONBuildingNet architecture, a multi-task neural network that extracts roof polygons and regresses offset vectors derived from satellite parallax clues. By learning the displacement between the roof and the ground footprint, it mathematically derives the absolute height.

Despite achieving high metric accuracy, LoD1 extrusion is fundamentally constrained by the "flat-roof assumption" [9]. By reducing vertical dynamics to a single scalar value, LoD1 models systematically fail to capture true 3D architectural manifolds, leading to "geometric collapse" where complex gables, dormers, and multi-level roof topographies are destroyed and simplified into shoebox representations.

Parametric Wireframe Reconstruction To capture mid-level visual abstractions, researchers shifted to wireframes, which define objects as a graph of junctions (corners) and edges with no faces, unlike meshes, see figure 2 for an example. Architectures like PBWR [12] attempt to regress 3D edges directly from aerial data by parameterizing each edge with a seven-parameter vector (including 3D midpoints, symbolic quadrants, and projected directional lengths). Conversely, BWFormer [10] adopts a "ground-up" strategy, projecting data into a 2D height map to detect corners before lifting them into 3D using edge-attention mechanisms. To leverage the power of generative modeling for these structural boundaries, recent frameworks like EdgeDiff [27] reconstruct building wireframes using diffusion models. EdgeDiff utilizes an Edge-Aware Denoising Module that explicitly parameterizes and denoises architectural edges by iteratively predicting their midpoints, geometric offsets, and confidence scores, and following that, they apply a clean-up algorithm similar to the Non Maximum Suppression algorithm to remove all redundant edges.

However, wireframes inherently face severe graph-connectivity challenges. A valid architectural wireframe must form closed loops to represent a watertight volume; yet, deep learning models frequently predict "floating" edges or disconnected junctions due to depth ambiguities [10]. Enforcing closed-loop connectivity from sparse aerial inputs is mathematically unstable, as the combinatorial space of potential connections often leads to conflicting gradients and entangled geometry [12].

Water-Tight Mesh Reconstruction Polygonal meshes provide the continuous surfaces necessary for physical simulation. Historically, deep learning frameworks have approached 3D mesh reconstruction by predicting implicit intermediate representations, such as continuous Occupancy Fields [28, 29] or Signed Distance Functions (SDFs) [30]. These continuous fields are subsequently discretized into explicit meshes using the

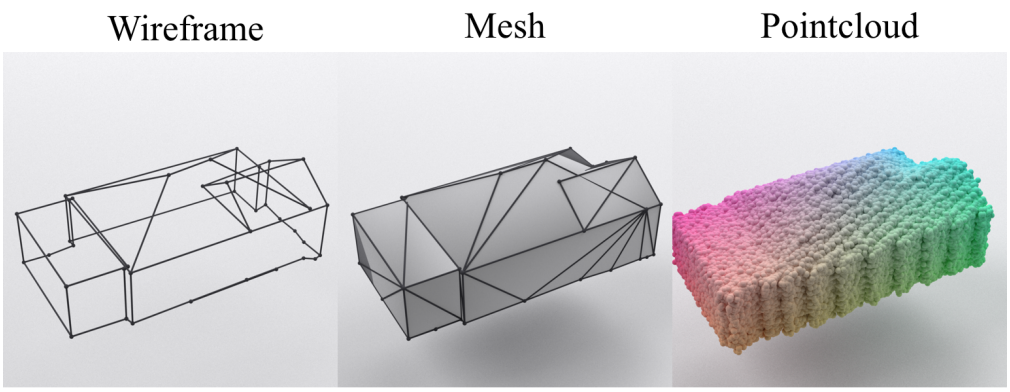


Figure 2: An illustration to show the difference between wireframe on the left, mesh in the middle, and point clouds on the right

classic Marching Cubes algorithm [31]. In the architectural domain, frameworks must handle extreme scale variations. Scale-ONet [32] addresses this by extracting occupancy fields conditioned on a scale factor, which is subsequently refined by Optim-Net to produce a watertight model.

Despite their ability to represent arbitrary topologies, these early implicit-to-explicit pipelines face severe geometric bottlenecks. GCN-based and implicit networks frequently cause "over-smoothing," rounding off the sharp, rectilinear corners characteristic of urban buildings [33]. Furthermore, because the standard Marching Cubes algorithm simply interpolates isovalues across a rigid volumetric grid, it introduces jagged artifacts at thin roof boundaries. To mitigate these artifacts, recent advancements in implicit modeling have replaced standard Marching Cubes with differentiable extraction techniques. Frameworks utilizing Deep Marching Tetrahedra (DMTet) [34] or FlexiCubes [35] dynamically optimize the underlying grid structures. State-of-the-art generative pipelines have rapidly integrated these modules; for instance, models such as GET3D [36] and Magic3D [37] leverage DMTet to synthesize highly detailed, high-resolution 3D meshes directly from 2D observations and diffusion priors. By allowing the network to adjust vertex positions alongside the implicit field during training, these methods successfully recover sharp geometric boundaries and significantly reduce topological noise. However, despite generating higher-fidelity surfaces, these advanced implicit methods still fundamentally produce unstructured "triangle soups." They lack meaningful architectural topological connectivity, which routinely complicates downstream CAD editing and physical simulation.

To overcome the severe limitations of unstructured dense triangulations, the field is undergoing a paradigm shift towards generating sparse, "artist-created" meshes that natively encapsulate true architectural topology, see figure 2 for an example. Inspired by Large Language Models, recent papers such as MeshGPT [38], MeshAnything [39], and MeshAnything V2 [40] model 3D geometry as a discrete sequence of vertices and faces. By utilizing autoregressive transformers, these methods directly predict highly structured, topologically clean, CAD-like meshes without relying on intermediate implicit fields. This autoregressive paradigm has been successfully adapted for the geospatial domain to reconstruct structured urban elements.

For example, Point2Building [33] reframes building reconstruction as a sequence-prediction task, utilizing an autoregressive transformer to iteratively generate explicit architectural vertices and faces. Sim-

ilarly, BuildAnyPoint [41] pushes this boundary by leveraging structured abstractions to directly predict topologically coherent 3D building models from diverse point clouds, ensuring the generated meshes mirror the logical structure of human-authored architectural designs.

B-rep and CAD Reverse Engineering The ultimate goal for architectural digital twins and BIM systems is the generation of B-rep CAD models. Unlike discrete meshes, B-rep defines 3D solids by explicitly modeling their boundaries. It achieves this by strictly decoupling geometry, the mathematical parameterization of underlying surfaces such as planes, cylinders, and Non-Uniform Rational B-Splines (NURBS), from topology, which dictates how these geometric elements connect via vertices, edges, and watertight faces. Because B-rep models are mathematically continuous and highly parameterized, they natively support complex downstream applications, including structural editing and physics-based simulations.

To achieve this level of parameterization, state-of-the-art (SOTA) frameworks typically approach B-rep generation as a 3D-to-3D reverse-engineering problem rather than a direct generative task. For example, Point2CAD [11] processes unorganized point sets by segmenting the inputs into distinct structural clusters, fitting geometric primitives via implicit neural representations, and systematically intersecting adjacent surfaces to recover sharp, continuous edges. Similarly, frameworks like Point2Cyl [13] focus on reverse-engineering specific extrusion primitives from point clouds to construct explicit cylindrical B-rep shapes.

Crucially, the current literature dictates that these complex B-rep structures cannot be reliably generated end-to-end directly from 2D pixels [11]. The direct prediction of the perfect joint distribution of geometry (surface parameters) and topology (graph connectivity) from a single image is mathematically intractable and highly unstable. Even a fractional angular error in a predicted surface normal vector prevents two parametric planes from properly intersecting, inevitably rendering the final CAD model non-watertight and topologically invalid.

The Intermediate Point Cloud Requirement The convergence of these geometric and topological limitations across all structured formats underscores the absolute necessity of utilizing unstructured 3D point clouds as an intermediate generative medium. Unlike meshes or B-rep models, point clouds are fundamentally unorganized sets of continuous (x, y, z) spatial coordinates. They completely lack rigid grid topologies, explicit surface boundaries, or connected edge graphs. Figure 2 illustrates an example of this raw, unstructured coordinate representation capturing an urban building's geometry.

While this lack of connectivity makes raw point clouds difficult for downstream simulation software to interpret natively, it provides an indispensable "metric-consistent scaffold" for deep learning architectures. By stripping away the mathematical burden of maintaining watertight surfaces or strict edge connectivity, generative models can freely model complex density variations. Utilizing this flexible, unconstrained representation allows the network to smoothly hallucinate the missing vertical volume and resolve complex architectural manifolds, such as multi-tiered roofs and sheer facades, without the rigid, premature topological commitments that cause direct structured prediction to collapse. Consequently, generating a dense, high-fidelity point cloud serves as the mathematically stable stepping stone, bridging the dimensional gap between flat pixels and the structured parametric surfaces required by smart cities.

2.1.3 The Generative Point Cloud Era

To bypass the severe topological constraints of structured representations and reconstruct unconstrained 3D architectures, the fundamental data representation had to shift toward unstructured 3D point clouds. By generating independent spatial coordinates, networks can freely model complex density variations and hallucinate missing vertical volumes without the mathematical burden of maintaining watertight surfaces or strict graph connectivity.

Because unordered point sets lack a rigid grid topology, their processing was revolutionized by the simultaneous adoption of Point Transformers and Denoising Diffusion Probabilistic Models (DDPMs). Diffusion models simulate a non-equilibrium thermodynamic process to generate data by iteratively reversing Gaussian noise, an approach translated to 3D point cloud synthesis by Luo and Hu [42]. Corley and Najafirad [43] brought this stochastic capability into the geospatial domain, demonstrating that diffusion models could generate statistically realistic structural textures. Processing these massive point sets was further accelerated by architectures like PTv3 [18], which bypassed quadratic attention bottlenecks by utilizing space-filling curves to serialize 3D coordinates.

Most critically, the community discovered that massive 2D LDMs harbor deep, pre-trained geometric priors. Xiong et al. [44] proved theoretically that spatial height and structural semantics are intrinsically entangled within these latent spaces. Frameworks like Marigold [16] demonstrated that LDMs could be fine-tuned to extract highly precise surface normals, while Viola et al. [45] showcased how sparse constraints could physically ground these diffusion hallucinations. Representing the current SOTA, architectures like BuildAnyPoint [41] actively merge these breakthroughs. By extracting high-fidelity geometric priors via latent diffusion and utilizing them to structurally condition the generative denoising of 3D point transformers, modern frameworks successfully bridge the dimensional gap, providing the exact unstructured geometric scaffold required for downstream pipelines to eventually fit parametric CAD models.

2.2 Deep Learning Architectures for 3D Point Clouds

Table 1 summarizes the chronological evolution of 3D point cloud architectures, highlighting the core mechanisms of each paradigm and their inherent limitations when applied to large-scale urban reconstruction tasks.

The reconstruction of large-scale urban geometry, specifically the generation of full 3D building point clouds from single, nadir-view aerial images, poses a formidable computational challenge. Unlike two-dimensional visual data neatly arranged on a regular Cartesian pixel grid, 3D point clouds exist as unordered, irregular, and highly sparse collections of continuous multidimensional coordinates. Historically, this unstructured nature precluded the direct application of standard CNNs, which mathematically require highly regular input formats to perform weight sharing and kernel optimizations.

Within the context of the proposed GeoPriorPC framework, generating massive, continuous vertical building facades requires an architectural backbone capable of processing millions of points while maintaining high-resolution fidelity. Crucially, it must command a massive global receptive field; to infer an occluded vertical facade from a single overhead image, the network must establish long-range structural dependencies, mathematically correlating the lowest visible ground point with the topmost roof edge. The evolution of 3D

Table 1: Chronological evolution of deep learning architectures for 3D point clouds, highlighting their mechanisms and limitations for large-scale geospatial reconstruction.

Era / Paradigm	Key Architectures	Core Mechanism	Limitations for Urban Reconstruction
Point-Based	PointNet [46], PointNet++ [47], DGCNN [48]	Shared MLPs, K-Nearest Neighbor (KNN) searches, Farthest Point Sampling (FPS).	$O(N^2)$ memory complexity and non-contiguous memory access make querying massive, continuous facades computationally intractable.
Voxel-Based	Minkowski Engine [49], PVCNN [50]	Sparse 3D Convolutions, mapping continuous float coordinates to discrete integer hash tables.	Quantization loss destroys sharp architectural edges; rigid convolutional kernels restrict the receptive field, breaking global structural context.
Transformer	PTv1 [51], OctFormer [52], PTv3 [18]	1D Serialization via space-filling curves, eliminating spatial querying for patch-based global attention.	*(Adopted by GeoPriorPC)* Operates with linear complexity, providing the massive global receptive field necessary to synthesize coherent, large-scale structural geometry.

point cloud architectures represents a continuous struggle to achieve this specific balance of computational scalability and geometric expressiveness.

2.2.1 The Point-Based models: MLPs and Spatial Querying

Early deep learning approaches relied on transforming point clouds into dense 3D volumetric grids, incurring an intractable cubic $O(N^3)$ memory penalty that wasted massive computational resources on empty space. The paradigm shifted with PointNet [46], which pioneered the direct processing of raw continuous coordinates using shared MLPs and symmetric global max-pooling to enforce permutation invariance. To capture finer architectural details, PointNet++ [47] introduced a hierarchical architecture utilizing Farthest Point Sampling (FPS) and continuous K-Nearest Neighbor (KNN) searches to aggregate local geometric neighborhoods.Subsequent architectures like DGCNN [48] constructed dynamic nearest-neighbor graphs in the feature space, while RandLA-Net [53] utilized random sampling to process larger scenes. However, the point-based era suffered from a fatal computational flaw: memory access patterns. Searching for continuous nearest neighbors in a high-dimensional space incurs an $O(N^2)$ complexity. Querying these neighbors forces the GPU to retrieve features from entirely random memory addresses, breaking contiguous cache fetching protocols and causing massive memory latency. To maintain computational feasibility, network designers were forced to keep their spatial groupings exceedingly small, severely restricting the network's receptive field and making the generation of large, continuous vertical facades virtually impossible.

2.2.2 The Voxel-Based models: Sparse Convolutions

Recognizing the intractable random memory access bottlenecks of KNN algorithms, researchers pivoted back to the spatial regularity of 3D grids, but with a critical optimization: sparse data structures. Frameworks utilizing Submanifold Sparse Convolutions [54] and the highly optimized Minkowski Engine [49]

mapped continuous coordinates into discrete integer hash tables. This allocated memory exclusively to occupied spatial coordinates, restoring the contiguous memory access patterns of traditional computer vision and enabling the processing of massive urban scenes. This era heavily influenced generative modeling, leading to hybrid models like PVCNN [50]. Despite these gains in efficiency, the voxel-based era introduced two fundamental geometric limitations. First, projecting continuous float coordinates into discrete grids via a floor operation introduces irreversible quantization loss. Fine architectural details, such as sharp roof parapets or window mullions, are mathematically collapsed into indistinguishable voxel centroids, destroying the crisp geometric edges required for structural modeling. Second, sparse 3D convolutions inherently possess a strictly localized receptive field dictated by their rigid kernel sizes (e.g., $3 \times 3 \times 3$). This sliding spatial window makes it exceedingly difficult for the network to route contextual information across massive physical distances, leading directly to the disjointed, "floating" wall geometries characteristic of early 3D diffusion priors.

2.2.3 Transformers: Serialization and Global Attention

To break free from localized convolutional biases, the field adapted the self-attention mechanisms of Transformers, allowing every point to dynamically compute its relevance to the entire structure. Early adaptations like Point Transformer v1 (PTv1) [51] and the Stratified Transformer [55] successfully implemented vector self-attention but remained bottlenecked by legacy spatial querying techniques, reviving the $O(N^2)$ memory constraints.

The definitive breakthrough for generative reconstruction was the abandonment of spatial querying in favor of 1D serialization. Models like OctFormer [52] proved that sorting algorithms could bypass spatial querying, but PTv3 [18] perfected this paradigm. PTv3 utilizes space-filling curves (such as Z-order and Hilbert curves) to traverse the multi-dimensional spatial volume, converting unstructured 3D points into a highly structured, 1D sequential token array.

By completely eliminating KNN searches and relative positional encodings, PTv3 operates with true linear complexity, running significantly faster than its predecessors while consuming vastly less GPU memory. More importantly, this extreme computational efficiency allows the architecture to expand its effective receptive field to an astonishing 1024 points per patch. By dynamically reshuffling the space-filling patterns between attention blocks, PTv3 guarantees that contextual information bleeds extensively across the entire 3D structure.

In the context of the proposed GeoPriorPC framework, replacing the legacy PVCNN backbone with a serialized DiPT [17] based on PTv3 provides the ultimate mathematical solution. When the diffusion process iteratively removes noise, the massive global receptive field allows tokens representing the rooftop to directly query tokens representing the ground-floor foundation in a single, unhindered attention layer. This long-range contextual awareness, completely free from quantization loss, guarantees that the generated 3D geometry maintains structural coherence across the entire building facade.

2.3 Data Modalities and Urban Benchmarks

The architectural successes of the structured 3D reconstruction frameworks discussed previously are inextricably linked to the data modalities upon which they are trained. A critical distinction in the current literature lies between models that perform 3D-to-3D processing versus those that perform true 2D-to-3D generation.

2.3.1 The Reliance on Active 3D Sensors

The vast majority of highly parameterized reconstruction networks, specifically those generating wireframes (e.g., BWFormer [10], PBWR [12]) and CAD B-reps (e.g., Point2CAD [11]), do not operate on 2D imagery. Instead, they require dense, pre-existing 3D point clouds as input, typically acquired via Airborne Laser Scanning (ALS) or Light Detection and Ranging (LiDAR). To train these models, the community relies on massive, city-scale 3D datasets such as RoofN3D [56], Building3D [57], and BuildingWorld [58], which provide hundreds of thousands of paired LiDAR point clouds and structured building models.

While LiDAR provides an unambiguous, metric-accurate geometric scaffold that makes fitting structured topologies mathematically tractable, it introduces a severe scalability bottleneck. Acquiring high-density aerial LiDAR is prohibitively expensive, logistically complex, and suffers from extremely low temporal revisit frequencies. Consequently, models that rely solely on 3D input modalities cannot be effectively deployed for rapid disaster response, real-time urban monitoring, or global-scale Digital Twin generation in developing regions where LiDAR data is unavailable.

2.3.2 Bridging the Modality Gap with Generative Benchmarks

Conversely, high-resolution 2D nadir aerial and satellite imagery is widely available, inexpensive, and updated at high temporal frequency. The grand challenge of the field, and the primary focus of this thesis, is to bridge the modality gap: utilizing the abundant 2D imagery to hallucinate the expensive 3D point clouds.

Historically, remote sensing datasets mapping 2D urban imagery to 3D structures were limited in scope. Foundational benchmarks like the ISPRS Vaihingen and Potsdam datasets [59] or the IEEE Data Fusion Contests [60] successfully paired orthophotos with DSMs, but these were strictly constrained to 2.5D raster regression. Similarly, massive global challenges like SpaceNet [61] focused heavily on 2D footprint extraction and LoD1 scalar height estimation. While massive, billion-point photogrammetric and LiDAR datasets exist—such as SensatUrban [62] and DublinCity [63], they are designed for scene-level semantic segmentation. They do not natively provide the isolated, instance-level pairs (a single building's image crop matched to a normalized 3D point cloud) required to train object-centric generative models.

It is only recently that the community has developed benchmarks capable of supporting true unconstrained 3D point cloud generation from single views. The AIM2PC dataset [1] represents a critical breakthrough in this domain. While AIM2PC is relatively small in scale, containing only 2,566 building instances, it is currently the *only* publicly available benchmark that strictly pairs high-resolution nadir RGB image crops with ground-truth, normalized 3D point clouds of diverse architectural topologies.

Despite its limited size, AIM2PC provides the exact empirical testbed necessary to train and evaluate 2D-to-3D generative diffusion models. Utilizing this unique dataset ensures that frameworks like GeoPri-

orPC can be evaluated not just on localized texture synthesis, but on their mathematical capacity to accurately resolve complex out-of-plane geometries from completely flattened 2D observations.

2.4 The Evolution of Generative Models

The transition from procedural modeling to data-driven 3D reconstruction requires generative frameworks capable of modeling highly complex, multi-modal distributions. Synthesizing urban structures from partial 2D observations dictates that a model must generate diverse, structurally coherent geometry without collapsing into average, smoothed representations. Over the past decade, deep generative modeling has undergone a paradigm shift to meet these exact requirements.

2.4.1 Pre-Diffusion Generative Baselines

Prior to the dominance of diffusion frameworks, the generative landscape was largely defined by two architectures: Generative Adversarial Networks (GANs) [64] and Variational Autoencoders (VAEs) [65]. GANs operate on a minimax game between a generator and a discriminator, producing highly realistic outputs. However, they are notoriously unstable to train and suffer from mode collapse, where the generator learns to produce only a limited subset of the target distribution. In the context of urban modeling, a mode-collapsed GAN might continuously generate standard flat-roofed houses while entirely failing to synthesize complex dormers or modern asymmetrical designs.

Conversely, VAEs explicitly optimize the Evidence Lower Bound (ELBO) to project data into a continuous latent space. While training is highly stable and ensures strong mode coverage, VAEs assume a Gaussian distribution for their latent variables. This often leads to over-smoothed, blurry generations, destroying the sharp, high-frequency boundaries (such as 90-degree facade corners) required for strict 3D architectural modeling.

2.4.2 The Mathematical Foundation of Diffusion

To overcome the instability of GANs and the blurriness of VAEs, the field turned to DDPMs. Originally conceptualized by Sohl-Dickstein et al. [66] using non-equilibrium thermodynamics, diffusion models were formalized into a highly effective generative framework by Ho et al. [67].

DDPMs operate via two Markov chains. The forward process gradually adds Gaussian noise to a real data sample x_0 over T timesteps, governed by a variance schedule β_t:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}) \tag{1}$$

Through the reparameterization trick, any noisy sample x_t at an arbitrary timestep t can be sampled directly in a closed form, as shown in Equation 2:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon \tag{2}$$

where $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and the added noise is $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$.

The generative reverse process trains a neural network (typically a U-Net) to reverse this degradation. Ho et al. [67] demonstrated that rather than predicting the noise parameters directly, the network $\varepsilon_\theta(x_t, t)$ can be trained to predict the added noise ε using a simplified Mean Squared Error (MSE) objective:

$$L_{simple} = \mathbb{E}_{x_0, \varepsilon \sim \mathcal{N}(0,\mathbf{I}), t} \left[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2 \right] \tag{3}$$

While DDPMs achieve state-of-the-art generation, their reliance on a full Markovian sampling chain makes them computationally slow during inference. This was heavily mitigated by Song et al. [68], who introduced DDIMs, allowing for deterministic, accelerated sampling trajectories by bypassing intermediate timesteps.

2.4.3 Latent Compression and 3D Adaptation

Despite algorithmic accelerations like DDIM, running the diffusion process directly in high-dimensional native data spaces, whether that involves millions of 2D pixels, dense 3D voxel grids, or massive unorganized point sets, remains computationally intractable. The definitive breakthrough that makes modern generative pipelines possible across all modalities was the introduction of LDMs by Rombach et al. [14].

Instead of diffusing raw, high-frequency data, LDMs utilize a pre-trained autoencoder to compress the input data x into a lower-dimensional, highly structural latent representation $z = \mathcal{E}(x)$. The diffusion process is then executed entirely within this efficient latent space, modifying the standard DDPM objective from Equation 3 to:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x), \varepsilon \sim \mathcal{N}(0,\mathbf{I}), t} \left[\|\varepsilon - \varepsilon_\theta(z_t, t)\|^2 \right] \tag{4}$$

Once the latent tensor z is fully denoised, the decoder $\mathcal{D}(z)$ projects it back into the original high-resolution space. Because this mathematical formulation completely separates perceptual data compression from semantic generative modeling, the LDM paradigm has been universally adopted across dimensions by simply swapping the underlying autoencoder architecture.

For 2D image synthesis, standard LDMs utilize spatial Autoencoder (AE) to map dense pixel grids into compressed spatial latents [14]. In the 3D domain, researchers have engineered highly specialized latent spaces to bypass the topological bottlenecks of rigid geometry. For generating unstructured 3D point clouds, frameworks like LION [69] employ hierarchical VAEs to project raw 3D coordinates into distinct vector-valued and point-structured latent distributions. Conversely, to generate continuous, watertight meshes, SOTA architectures rely on implicit representations. For example, XCube [70] compresses highly parameterized 3D grid data into a Sparse Voxel Hierarchy prior to diffusion, while frameworks like GetMesh [71] project structured geometric features onto latent triplanes, from which final topological meshes are extracted via differentiable tetrahedral algorithms.

Within the proposed GeoPriorPC framework, the first stage directly capitalizes on this decoupled efficiency. By fine-tuning a foundational LDM originally pre-trained on pixel-space AEs, the framework successfully adapts the model to extract a precise, structural intermediate prior, specifically, compressed latent normal maps, from top-down imagery. This extracts critical out-of-plane geometries without incurring the massive computational overhead of decoding back into high-resolution pixel space.

Simultaneously, the mathematical principles of DDPMs were adapted to operate directly on the target 3D representation. Luo and Hu [42] formalized the first 3D diffusion process, treating unstructured point coordinates as particles governed by a thermodynamic heat bath. This was subsequently expanded into highly conditional pipelines by architectures like Point-E [72], which demonstrated that 3D point cloud diffusion could be heavily guided by localized image features.

To operationalize these theoretical principles, the proposed GeoPriorPC framework explicitly utilizes a 2D LDM for the critical task of generating the intermediate geometric prior. Rather than attempting to learn spatial structures from scratch, Stage 1 fine-tunes a foundational, pre-trained 2D LDM specifically to predict oblique surface normal maps from top-down nadir imagery. This strategic adaptation directly harnesses the vast generative power and robust generalization capabilities embedded within these massive vision models, enabling the network to reliably hallucinate occluded out-of-plane geometries based on its pre-existing "world knowledge." Ultimately, by merging the extreme computational efficiency of this latent 2D diffusion (Stage 1) with the structural adaptability of unconstrained 3D point diffusion (Stage 2), the foundation is laid for the dual-conditioned architecture proposed in this thesis.

2.5 Generative Priors for 3D Geometry

The fundamental limitation of earlier 3D reconstruction models was their reliance on geometric priors learned exclusively from small-scale, domain-specific geospatial datasets. Because these datasets lack the diversity required to model all possible architectural variations, the resulting networks struggled to generalize. However, the advent of massive, internet-scale LDMs—such as Stable Diffusion [14]—has introduced a new paradigm: leveraging foundational 2D models for 3D physical understanding.

2.5.1 The Implicit 3D Knowledge of Foundation Models

Although text-to-image LDMs are trained purely on 2D pixel grids to minimize a denoising objective, recent probing studies have revealed that their latent spaces implicitly encode robust physical and geometric properties. To successfully hallucinate complex scenes under varying lighting conditions, the U-Net backbone of an LDM is forced to internally model depth, occlusion, surface normals, and perspective [15]. This phenomenon implies that foundational LDMs possess a powerful, generalized "world prior", a mathematical understanding of how 3D manifolds project into 2D space.

This emergent 3D awareness is a direct mathematical consequence of the model's massive training scale. To efficiently minimize reconstruction loss across billions of 2D images, the latent space discovers true physical factors of variation as a form of optimal data compression. Empirical investigations into the internal activations of frozen diffusion U-Nets confirm that they maintain dense, multi-view consistent structural representations rather than superficial 2D textures [73]. Consequently, multiple studies have demonstrated that these internal representations inherently capture depth, geometry, and viewpoint correspondence [16, 74], effectively operating as implicit rendering engines. The paradigm of modern 3D computer vision has therefore shifted: rather than attempting to learn 3D geometry from scratch using limited, domain-specific geospatial datasets, the current imperative is to unlock and harness this generalized 2D prior for explicit 3D generation.

2.5.2 Extracting the Prior: Optimization vs. Feed-Forward

The challenge in modern 3D computer vision is extracting this implicit 2D prior and physically grounding it into a 3D representation. The literature currently divides this extraction process into two distinct paradigms:

1. **Optimization-Based Extraction:** Frameworks like DreamFusion [75] introduced Score Distillation Sampling (SDS). SDS treats a pre-trained 2D diffusion model as a critic, iteratively optimizing a 3D representation (such as a Neural Radiance Field or point cloud) so that its 2D renderings look realistic to the diffusion model at any angle. Subsequent frameworks such as Magic3D [37] and ProlificDreamer [76] built upon this foundation by introducing two-stage optimizations and Variational Score Distillation (VSD), respectively, significantly improving geometric fidelity and mitigating over-smoothing artifacts. However, despite these advancements, the iterative optimization required for every single object makes this paradigm computationally intractable for large-scale urban reconstruction.

2. **Feed-Forward Fine-Tuning:** To bypass the immense computational cost of SDS, researchers developed methods to fine-tune the LDM's U-Net to directly output explicit geometric maps in a single feed-forward pass. Architectures like ControlNet [77] and T2I-Adapter [78] proved that LDMs could be locked and augmented with trainable feature pathways to map structural inputs (like Canny edges) to generated images. Conversely, moving from structure-to-image to image-to-structure, the Marigold framework [16] and GeoWizard [79] demonstrated that the pre-trained diffusion backbone could be fine-tuned to predict highly accurate, dense geometric representations, such as depth and surface normals, directly from RGB inputs. Furthermore, models like Wonder3D [80] extend this feed-forward paradigm to generate cross-domain, multi-view normal maps in a single inference step, directly providing the geometric scaffolds necessary for deterministic 3D reconstruction.

2.6 The Gap in the Literature

While feed-forward models like Marigold successfully extract deterministic surface normal maps, a critical gap remains in the geospatial literature regarding the final usability of these outputs. A 2D normal map, while structurally informative, is fundamentally an optical illusion constrained to the image plane. It lacks explicit spatial coordinates, true volumetric density, and absolute scale. Consequently, these intermediate representations cannot be natively ingested by urban simulation software, BIM pipelines, or Digital Twin ecosystems that strictly require unstructured 3D coordinates for spatial analysis and downstream B-rep reverse-engineering.

Conversely, existing single-stage 3D generative baselines that attempt to map nadir aerial pixels directly to unstructured point clouds (such as AIM2PC [1]) suffer from severe geometric ambiguity. Because these networks lack an explicit structural anchor, they are forced to blindly hallucinate the occluded vertical dimensions. This routinely results in collapsed facades, "pancaked" out-of-plane variance, and blurred topological features that destroy the architectural integrity of the generation.

Therefore, the literature currently exhibits a severe methodological disconnect. There is no cohesive framework designed to effectively bridge these two powerful generative domains: the generalized struc-

tural reasoning of foundational 2D latent diffusion, and the unconstrained geometric flexibility of 3D point cloud diffusion. The primary mathematical bottleneck preventing this union has been the lack of a robust fusion mechanism capable of systematically injecting dense 2D latent priors into the permutation-invariant, unordered space of 3D coordinates without severely compromising computational scalability.

Addressing this specific architectural gap is the primary motivation for the GeoPriorPC framework proposed in this thesis. By purposefully decoupling the generative process, GeoPriorPC extracts the structural foundation prior via 2D latent diffusion and subsequently injects it into a highly scalable, serialized 3D Diffusion Point Transformer (DiPT). Through the introduction of a novel dual-conditioning mechanism, utilizing both global adaptive layer normalization and localized cross-attention, this thesis provides the missing mathematical bridge, enabling the synthesis of highly accurate, unstructured building geometry directly guided by foundational world knowledge.

3 GeoPriorPC: A two-stage diffusion model for 3D building reconstruction

This chapter details the methodology and architectural design of GeoPriorPC, a novel two-stage generative framework developed to reconstruct 3D building point clouds from single nadir-view aerial images. The chapter begins by formalizing the problem space and reviewing the mathematical foundations of both 2D latent diffusion and 3D point cloud diffusion. Subsequently, it provides a comprehensive breakdown of the two distinct stages of the proposed pipeline: the extraction of geometric priors using foundational models, and the geometry-conditioned 3D generation process. Finally, this chapter documents the systems engineering, training dynamics, and empirical architectural explorations undertaken to finalize the model design.

3.1 Overview of the Architecture

Reconstructing accurate 3D urban geometry from a single top-down (nadir) view is a fundamentally ill-posed problem. The orthogonal projection captures roof structures effectively but completely occludes vertical geometry, requiring the model to infer the missing out-of-plane dimensions. Traditional single-stage generative methods struggle to fuse raw 2D image features with the plausible structural priors necessary to synthesize complex vertical facades and accurate roof proportions.

To resolve this inherent ambiguity, the GeoPriorPC framework decouples the image-to-point cloud task into a two-stage generative pipeline. Rather than forcing a single network to simultaneously guess the hidden geometry and generate the 3D coordinates, the architecture leverages the generalized world knowledge embedded in foundational models as an explicit intermediate step.

The framework operates through two distinct but interconnected stages:

1. **Stage 1: Geometric Prior Learning (Image-to-Image):** In the first stage, the framework adapts a pre-trained LDM to predict the surface normals of the corresponding oblique view from the nadir input. Unlike traditional methods that target RGB textures, this stage focuses on generating stable and sharp geometric maps. Crucially, the output is not decoded back into pixel space; it is retained as a compressed latent representation to serve as a robust, feature-rich spatial prior.

2. **Stage 2: Geometry-Conditioned 3D Generation (Image-to-Point):** The second stage maps the 2D spatial features and the learned structural prior into the 3D domain. Nadir image features are projected onto a noisy 3D point cloud, which is iteratively denoised using a DiPT backbone. A dual-conditioning mechanism explicitly guides the 3D denoising process, fusing the latent geometric prior globally via adaLN-Zero layers and locally through cross-attention to ensure fine-grained spatial alignment.

3.2 Mathematical Foundations of Diffusion Models

The GeoPriorPC framework relies on the principles of DDPM, applied across both 2D latent image spaces and unstructured 3D point sets. Diffusion models operate via a two-step Markov chain: a forward diffusion process that progressively corrupts the input data with Gaussian noise, and a learned reverse process that iteratively reconstructs the original data distribution from pure noise.

3.2.1 2D Latent Diffusion Models

Traditional DDPMs operate directly in the high-dimensional pixel space. While effective, calculating self-attention and convolutions across dense spatial grids (e.g., $512 \times 512 \times 3$ pixels) is computationally exorbitant. To mitigate this bottleneck, LDMs decouple the generative process into two distinct components: a perceptual compression stage via an Autoencoder, and a semantic generation stage via the diffusion model.

The Autoencoder Architecture The compression stage relies on an off-the-shelf, pre-trained Autoencoder consisting of an Encoder $\mathcal{E}$ and a Decoder $\mathcal{D}$.

- The Encoder ($\mathcal{E}$): The encoder compresses a high-resolution target image $x \in \mathbb{R}^{H \times W \times 3}$ into a lower-dimensional latent representation $z_0 = \mathcal{E}(x)$, where $z_0 \in \mathbb{R}^{h \times w \times c}$. This latent space discards imperceptible high-frequency details, focusing instead on the core semantic and structural composition of the image.
- The Decoder ($\mathcal{D}$): The decoder performs the inverse operation, mapping the latent representation back to the pixel space, such that $\tilde{x} = \mathcal{D}(z_0) \approx x$.

Pre-training and Regularization Dynamics Crucially, this Autoencoder is integrated entirely as a frozen, off-the-shelf module without any custom pre-training or parameter updates. Following the standard Latent Diffusion framework, the weights were initially pre-trained on the massive Open Images dataset [81] and subsequently fine-tuned on the LAION dataset [82] to maximize reconstruction fidelity.

The architecture is optimized using a hybrid combination of a pixel-wise reconstruction loss, a patch-based adversarial constraint, and a perceptual loss (such as LPIPS) to ensure the decoded outputs remain visually crisp. Additionally, a mild Kullback-Leibler (KL) regularization term is applied during training. Rather than forcing a strict probabilistic parametric mapping as seen in classic variational models, this penalty is heavily weighted down to act strictly as a loose constraint, bounding the latent space variance to ensure a continuous and well-behaved distribution that the diffusion model can stably traverse.

The Diffusion Process in Latent Space Once the Autoencoder weights are frozen, the forward diffusion process progressively adds Gaussian noise strictly to the clean latent z_0 over T timesteps, producing a sequence of noisy latents $z_1, \ldots, z_T$.

During the reverse generative process, the U-Net (ε_θ) is trained to predict the added noise ε. Conditioned on an external spatial signal c, the model is optimized using the DDPM objective:

$$L_{LDM} = \mathbb{E}_{z_0, \varepsilon, t} \left[||\varepsilon - \varepsilon_\theta(z_t, t, c)||_2^2 \right] \tag{5}$$

During inference in GeoPriorPC, we leverage the pre-trained Autoencoder encoder to extract latent features from the nadir inputs. However, to preserve dense geometric features for the 3D generation stage and reduce computational overhead, we explicitly bypass the Autoencoder decoder, utilizing the denoised latent $\hat{z}_0$ directly as our intermediate condition.

3.2.2 Point Cloud Diffusion: Operating in the Direct Coordinate Space

While the 2D stage requires an AE to manage the density of pixel grids, the final spatial reconstruction extends the diffusion framework to operate directly on unstructured 3D sets. Let $P_0 = P$ denote the ground-truth 3D building coordinates, where $P \in \mathbb{R}^{N_p \times 3}$ and N_p is the total number of points.

Contrasting 3D vs. 2D Diffusion Spaces Unlike dense 2D images, 3D point clouds are inherently sparse representations. Because a building shape can be effectively represented by $N_p = 10,000$ points, there is no strict need for a AE to compress the data into a latent space. Therefore, the 3D diffusion process operates directly in the explicit Cartesian coordinate space.

The forward point diffusion process systematically perturbs these 3D coordinates over T timesteps. The injection of noise is governed by a predetermined variance schedule β_t, which dictates the scale of the Gaussian noise added at each step. The transition probability from state P_{t-1} to P_t is formally defined as:

$$q(P_t|P_{t-1}) = \mathcal{N}(P_t; \sqrt{1-\beta_t}P_{t-1}, \beta_t I) \tag{6}$$

As $t \to T$, the point cloud P_T loses all structural information and conforms to an isotropic Gaussian distribution $P_T \sim \mathcal{N}(0, I)$. To maintain numerical stability across the DiPT backbone, the coordinates are globally normalized into a unit sphere. We employ a Centered Denoising Probabilistic Model (CDPM) [83] strategy to maintain a zero-mean point cloud across both the forward and reverse operations.

The CDPM is a diffusion variant designed to overcome the uncontrollable center deviations that degrade single-image 3D reconstruction in standard models. Its core mechanism restricts the added noise, the predicted noise, and the sampled 3D point cloud into a constrained subspace where their gravity centers consistently coincide with the origin. By strictly maintaining a zero-mean distribution—centralizing the data and noise at every step, CDPM prevents the network from wasting capacity on tracking the object's shifting center. Crucially, this stable center acts as a fixed anchor; when the point cloud is back-projected onto the 2D image, it ensures the extracted local features remain perfectly aligned and consistent throughout the denoising process, though this stability comes at the cost of slightly reduced generative diversity.

The reverse denoising process iteratively recovers the clean architectural geometry P_0 directly from the noisy distribution P_T. This 3D coordinate synthesis is explicitly guided by our dual-conditioning mechanism, which fuses the explicit 2D spatial context extracted from the nadir inputs with the intermediate latent geometric prior predicted by the LDM.

3.3 Problem Formulation and Input Representation

The primary objective of the GeoPriorPC framework is to reconstruct a comprehensive 3D building structure from a strictly top-down 2D observation.

3.3.1 Input Space Definition

The model takes a multimodal set of 2D representations as input. Let the primary input be a single nadir-view RGB aerial image $I \in \mathbb{R}^{H \times W \times 3}$. To provide explicit spatial boundaries and semantic localization, the

framework is concurrently fed a building footprint mask $M_b \in \{0,1\}^{H\times W\times 1}$ and a corresponding Sobel edge map $E \in \mathbb{R}^{H\times W\times 1}$.

Spatial projection between the 2D plane and the 3D domain is governed by the associated camera parameters $C = \{K, R, t\}$. Given that the underlying dataset consists of standardized aerial acquisitions, we operate under the assumption that the intrinsic matrix K and the translation vector t remain constant across all samples. Consequently, the rotation matrix R serves as the primary spatial variable facilitating 2D-to-3D geometric alignment.

3.3.2 Target Output Space

The target output of the generative pipeline is a reconstructed 3D building point cloud denoted as $P \in \mathbb{R}^{N_p\times 3}$, where N_p represents the total number of points (standardized to $N_p = 10,000$ in our methodology).

It is important to note that the target point cloud P is defined within a normalized coordinate space, specifically scaled into a unit sphere. The generative objective focuses on recovering the relative 3D structural topology, architectural complexity, and geometric proportions of the building, rather than its absolute metric dimensions. This normalization strategy ensures mathematical stability during the diffusion process and is consistent with standard generative 3D reconstruction benchmarks.

3.3.3 The Formulation of the Geometric Prior

While the inputs I, M_b, and E successfully provide foundational 2D spatial context and boundary constraints, they are mathematically insufficient for determining out-of-plane (vertical) geometry. A single mapping from a nadir pixel (x, y) to a 3D coordinate (x, y, z) cannot natively resolve co-planar vertical points, such as facades.

Therefore, the problem formulation requires the generation of an intermediate variable to resolve this inherent vertical ambiguity. Rather than predicting high-resolution pixel-space depth or normals, the framework formulates this intermediate condition as a compressed latent surface normal representation $\hat{z}_0 \in \mathbb{R}^{h\times w\times c}$. Here, $h = H/8$, $w = W/8$, and $c = 4$ represent the spatial dimensions downsampled by the Autoencoder. This latent prior explicitly conditions the final generative mapping from the 2D observation space to the 3D coordinate space P.

3.4 Stage 1: Geometric Prior Extraction and Data Engineering

To successfully condition the 3D generation process, the framework first extracts a reliable geometric prior from the nadir observation. This section details the data engineering pipeline required to render ground-truth oblique normals, the optimization of the latent encoding process, and the training dynamics of the fine-tuned LDM. It concludes with an empirical analysis of multi-view versus single-view prior extraction.

3.4.1 Data Preprocessing and Oblique Normal Rendering

The fine-tuning of the LDM requires paired tuples of nadir RGB images and their corresponding ground-truth normal maps (I, N_{GT}). Because the foundational AIM2PC [1] dataset provides only raw 3D point clouds,

the explicit geometric normals must be derived programmatically.

First, dense 3D surface meshes are reconstructed from the ground-truth point clouds. Point normals are estimated using Open3D [84], and the discrete points are converted into continuous watertight surfaces via Poisson Surface Reconstruction[85].

Once the meshes are generated, oblique normal maps are rendered on-the-fly using the PyTorch3D library[86]. To maintain a consistent spatial relationship with the top-down aerial image I, the rendering pipeline employs an `OrthographicCamera` setup. The camera is parameterized with a focal length of 0.8 and positioned using a look-at view transform set to a distance of 2.0, an elevation angle of 40°, and a base azimuth of 180°. This specific 40° elevation is critical, as it perfectly simulates an oblique aerial perspective, explicitly exposing the vertical facades that are hidden in the nadir view.

The rendered surface normals, originally represented as 3D directional vectors, are then normalized into the $[-1,1]$ continuous range to form the channels of the target image N_{GT}. At the same time, the input image I is subtracted from a background mask to remove all background details.

To ensure the model learns robust geometric invariance and avoids overfitting, rigorous data augmentation is applied during this generation phase. Random rotations and horizontal flipping are applied. Crucially, to maintain perfect 2D-to-3D spatial correspondence, the exact same rotation and flip transformations are applied to both the input nadir image and the 3D mesh *prior* to the PyTorch3D rasterization step. You can see some data samples created from this pipeline in figure 3.

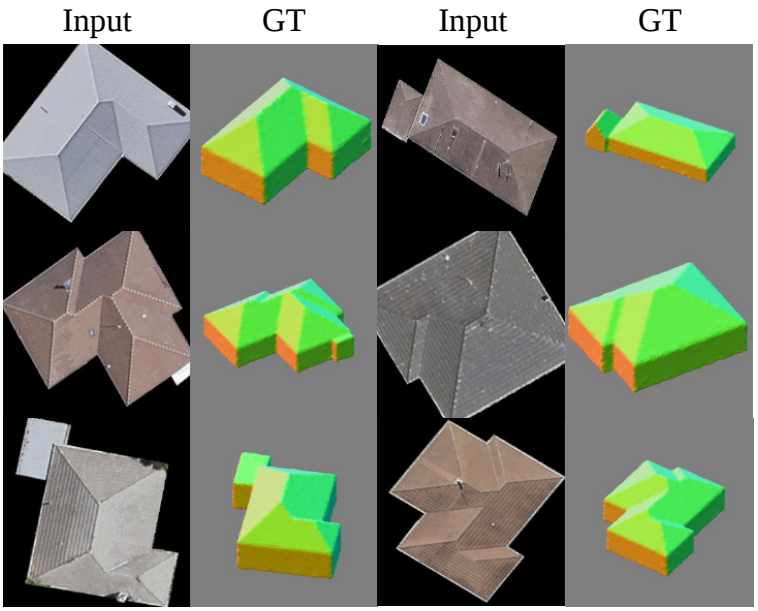


Figure 3: Samples of the generated dataset for LDM training. Each sample contains an RGB image with mask subtraction and a target normal oblique image. All processing and rendering run on the fly during training.

3.4.2 Latent Space Encoding and Computational Optimization

Following the Marigold framework [16], both the augmented input aerial image I and the target normal map N_{GT} are mapped into a compressed latent space using a pre-trained AE encoder $\mathcal{E}$, producing latents

$z_I = \mathcal{E}(I)$ and $z_0 = \mathcal{E}(N_{GT})$.

During standard diffusion training, encoding high-resolution images dynamically at every step introduces a severe computational bottleneck and excessive VRAM consumption. To optimize the training pipeline, the AE weights are strictly frozen. Furthermore, an offline pre-computation strategy is employed: all input images and target normal maps are resized to 256×256, encoded into their respective latents (z_I and z_0), and saved to disk prior to the initiation of the diffusion training loop. By loading the pre-computed latents directly, the framework maximizes batch throughput and dedicates the entire GPU memory budget to the U-Net denoising process.

3.4.3 Latent Diffusion Fine-Tuning Dynamics

To learn the robust geometric mapping from a nadir RGB image to an oblique surface normal, the LDM undergoes a fine-tuning process. This phase is divided into two distinct operational modes: the forward training process, governed by the standard DDPM objective, and the accelerated reverse inference process, governed by non-Markovian DDIM sampling.

Forward Process and Training Objective During training, Gaussian noise ε is progressively added to the clean normal latent z_0 to generate a noisy latent z_t at timestep t. To provide the necessary 2D spatial context for the generation, we then concatenate the clean image latent z_I with the noisy normal latent z_t along the channel dimension. Forming a combined input tensor of size $h \times w \times 8$, this concatenation serves as an explicit, pixel-aligned condition for the network. Because the pretrained U-Net natively supports only 4 input channels, we extended the dimensions of the first CNN layer by duplicating the pre-trained weights. We then divided all weight values in this modified layer by 2, ensuring the layer's output activation range remained consistent with the original network.

To enhance the model's spatial adaptability, the training procedure is structured in two distinct phases regarding image dimensions. Initially, the input nadir images and corresponding normal maps are forcibly resized to a square resolution of 256×256, explicitly ignoring the original aspect ratios of the raw aerial acquisitions. Following this 1:1 aspect ratio training, the model undergoes a second fine-tuning phase where it is trained on data with varying aspect ratios. This curriculum strategy ensures the LDM inherently learns to process and understand different spatial proportions.

The diffusion U-Net ε_θ, initialized with Stable Diffusion 2.0 weights, processes the concatenated input to predict the added noise. By directly learning the noise residual, the model acts as a robust denoiser. The optimization objective is formulated as the standard mean squared error between the true noise $\varepsilon \sim \mathcal{N}(0, I)$ and the predicted noise:

$$L_{prior} = \mathbb{E}_{z_0, \varepsilon \sim \mathcal{N}(0,I), t} \left[||\varepsilon - \varepsilon_\theta(z_t \oplus z_I, t)||_2^2 \right] \tag{7}$$

The model was trained extensively using a batch size of 32 over 250 epochs. This hyperparameter configuration ensures stable gradient updates while providing the network with sufficient structural diversity per batch to learn the complex mapping from nadir textures to oblique geometries. The overall training workflow, including the forward noise injection and conditional concatenation, is illustrated in Figure 4.

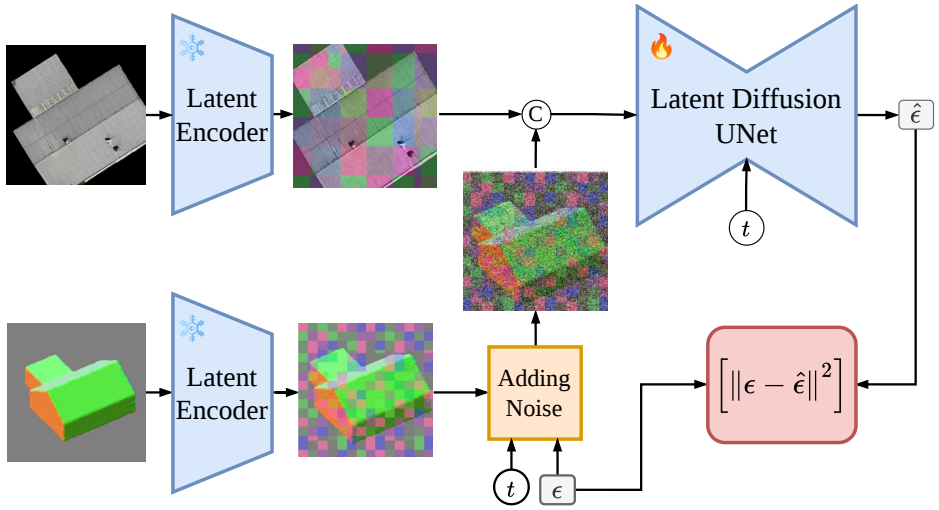


Figure 4: The forward training process of the geometric prior LDM. Gaussian noise is added to the ground-truth normal latent, concatenated with the encoded RGB image latent, and passed through the U-Net to compute the noise prediction loss.

Non-Markovian Inference and Sampling While the model is optimized using the Markovian DDPM objective during training, strictly following this trajectory during inference is computationally expensive due to the high number of sequential timesteps required. To accelerate the extraction of the geometric prior during test-time, the framework transitions to DDIM for sampling.

Crucially, unlike the initial training phase, the input nadir image is *not* resized to a square during inference. The original aspect ratio of the image is explicitly preserved and fed directly into the model. Because the LDM was taught to comprehend varying aspect ratios during its two-stage training curriculum, the model processes these arbitrary dimensions seamlessly without generating spatial distortions.

During inference, we initialize a completely noisy target latent $z_T \sim \mathcal{N}(0, I)$. The model iteratively denoises it over time using the fine-tuned U-Net, concatenating it with z_I at each step. By leveraging the non-Markovian formulation of DDIM, Stage 1 samples the oblique normal latent $\hat{z}_0$ efficiently using only $T_1 = 50$ steps, a linear noise schedule, $\eta = 0$, and no classifier-free guidance.

Upon completing the reverse sampling process, we obtain the clean predicted normal map in the latent space. Rather than passing the final denoised latent $\hat{z}_0$ through the AE decoder, we use it directly as the intermediate spatial condition for the subsequent 3D generation stage. Bypassing the decoder preserves the dense, high-frequency structural semantics encoded in the latent space and significantly reduces the pipeline's computational overhead by not requiring a large pretrained image decoder. The complete inference trajectory is depicted in Figure 5.

The primary hyperparameters for this stage are summarized in Table 2.

3.4.4 Architectural Explorations

Aspect Ratio Conditioning We investigated whether explicitly conditioning the U-Net on the input image's aspect ratio would further improve the extraction of the geometric prior. To test this hypothesis, we designed a custom aspect ratio embedding module consisting of two MLP layers.

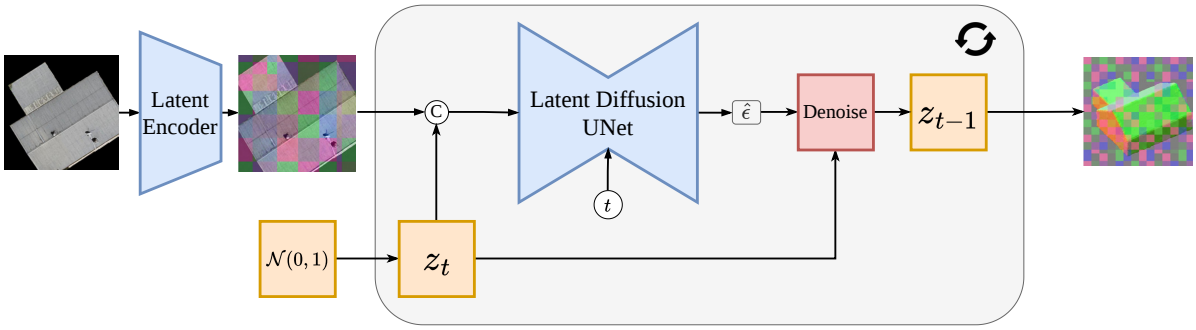


Figure 5: The accelerated inference process using DDIM sampling. The model denoises pure Gaussian noise over 50 steps, heavily conditioned on the nadir image latent, to output the final compressed geometric prior $\hat{z}_0$.

Table 2: Hyperparameters for Data Rendering and Stage 1 LDM Fine-Tuning.

Parameter	**Value**
Data Rendering (PyTorch3D)	
Camera Distance	2.0
Elevation Angle	40°
Base Azimuth Angle	180°
Focal Length	0.8
Camera Type	Orthographic
Vector Normalization	$[-1, 1]$
Latent Diffusion Training	
Base Architecture	Stable Diffusion 2.0
Initial Input Resolution	256×256 (Stage 1)
Batch Size	32
Training Epochs	250
AE Encoder Weights	Frozen

This module was engineered to project a scalar aspect ratio value into a high-dimensional embedding matching the exact size of the standard timestep embedding. This resultant aspect ratio embedding was then additively combined with the time embedding before being injected into the U-Net layers via cross-attention. To ensure that this architectural modification did not destabilize the pre-trained weights of the Stable Diffusion backbone at the start of training, the MLP layers were initialized entirely with zeros, see figure 6 for visual explanation.

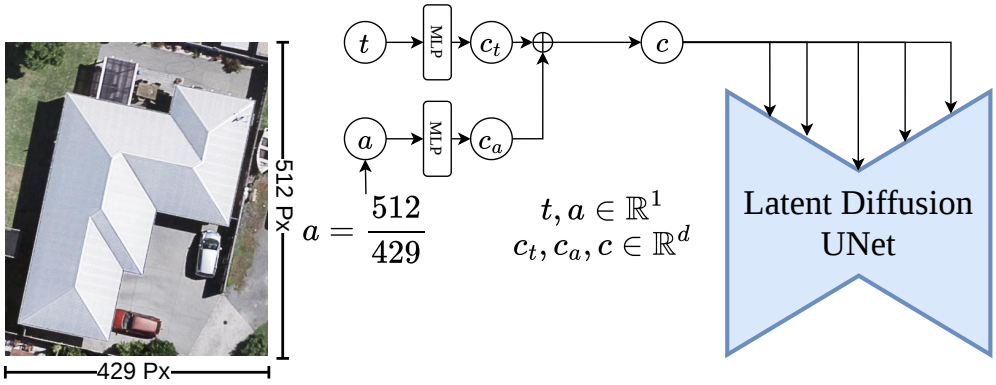


Figure 6: Explanation of aspect ratio injection. We are using an MLP to create an embedding with the same size as the time embedding, which is then added together and fed to the model via cross-attention.

Empirical evaluation revealed that while this explicit aspect ratio conditioning provided a marginal improvement in the Stage 1 normal map prediction metrics, it yielded zero observable downstream improvement in the final 3D point cloud generation (Stage 2). Because the two-stage training curriculum already allowed the network to implicitly handle varying aspect ratios, the explicit module was deemed redundant for the final pipeline. A detailed quantitative breakdown of this ablation will be discussed in Section 4.5.5.

Single-View vs. Multi-View Priors During the developmental phase of the framework, we hypothesized that providing the 3D generation stage with multiple geometric priors from different perspectives would yield a more comprehensive structural understanding of the building.

To test this architectural variant, the PyTorch3D rendering pipeline was expanded to generate four distinct views by varying the yaw angle $\{0°, 90°, 180°, 270°\}$ relative to the base azimuth. The LDM was subsequently modified to predict these multiple views to serve as a composite condition for the 3D generative backbone. An example of the input and output from the model can be seen in figure 7.

However, as will be detailed in Section 4.x (Ablation Studies), empirical evaluation revealed inherent challenges with cross-view geometric consistency when utilizing independent diffusion generations. Consequently, to maintain strict geometric coherence and reduce computational overhead, the final GeoPrior-Diff framework was streamlined to utilize a single-view geometric prior, specifically utilizing the primary yaw angle (0°).

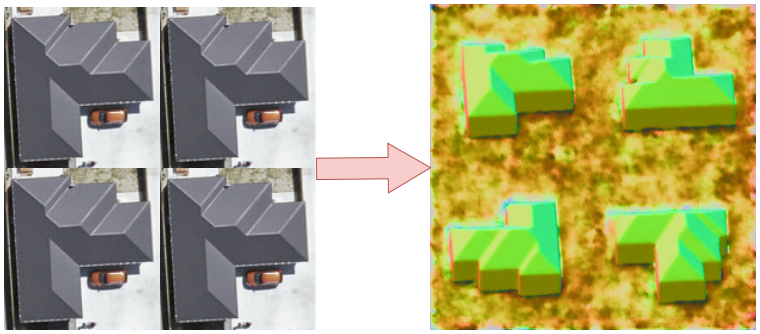


Figure 7: An input and output example from the multiview LDM. The model is now expecting to receive the same image duplicated 4 times, and for each duplication, the model is predicting a different oblique normal view.

3.5 Stage 2: Geometry-Conditioned 3D Point Generation

While Stage 1 successfully extracts an intermediate geometric prior from the 2D observation space, Stage 2 is responsible for synthesizing the final 3D building structure. This section details the core generative backbone of GeoPriorPC, which extends the principles of point cloud diffusion by integrating a custom DiPT architecture. We explicitly detail the multi-modal feature projection module, the adaptation of the PTv3 backbone, and the novel dual-conditioning mechanism designed to fuse the Stage 1 normal latent into the 3D denoising process.

3.5.1 Multi-Modal 2D Feature Extraction and Projection

To ensure the generated point cloud adheres strictly to the semantic boundaries and textural details of the original observation, the 3D diffusion process must be tightly coupled with the 2D input space.

First, we construct a comprehensive 2D spatial context by concatenating the nadir RGB image I, the building footprint mask M_b, and the Sobel edge map E along the channel dimension. A pre-trained convolutional encoder, specifically Vision Transformer (ViT), processes this combined input to extract dense, hierarchical spatial features, denoted as F_{2D}.

To map these features into the 3D domain, we employ a differentiable projection mechanism. At any given diffusion timestep t, the noisy 3D point cloud P_t is back-projected onto the 2D image plane using the fixed camera parameters $C = \{K, R, t\}$. Because the continuous 3D coordinate projections rarely align perfectly with the discrete 2D pixel grid, we utilize bilinear interpolation to sample the feature map. For each 3D point $p_i \in P_t$, its corresponding projected 2D location (u_i, v_i) is used to extract a localized feature vector f_i^{2D} from F_{2D}. These localized 2D features are subsequently concatenated with the 3D point coordinates and the current timestep embedding to initialize the input point features for the transformer backbone.

3.5.2 The Diffusion Point Transformer (DiPT) Backbone

Standard 3D convolutional architectures and early point-based networks (e.g., PointNet++) struggle to scale efficiently for dense diffusion tasks. To process the 10,000-point clouds generated in our framework, we employ a DiPT backbone. DiPT adapts the highly efficient PTv3 architecture for generative denoising.

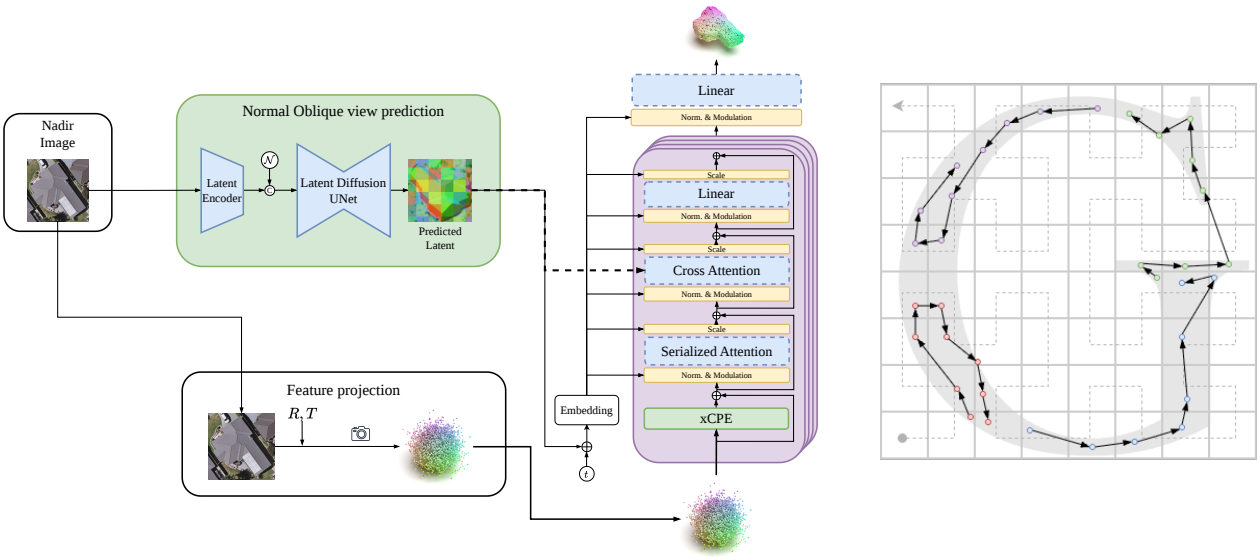


Figure 8: The proposed GeoPriorPC main. On the left, the model architecture is based on DiPT, while on the right, an illustration of the serialized process used in serialized attention.

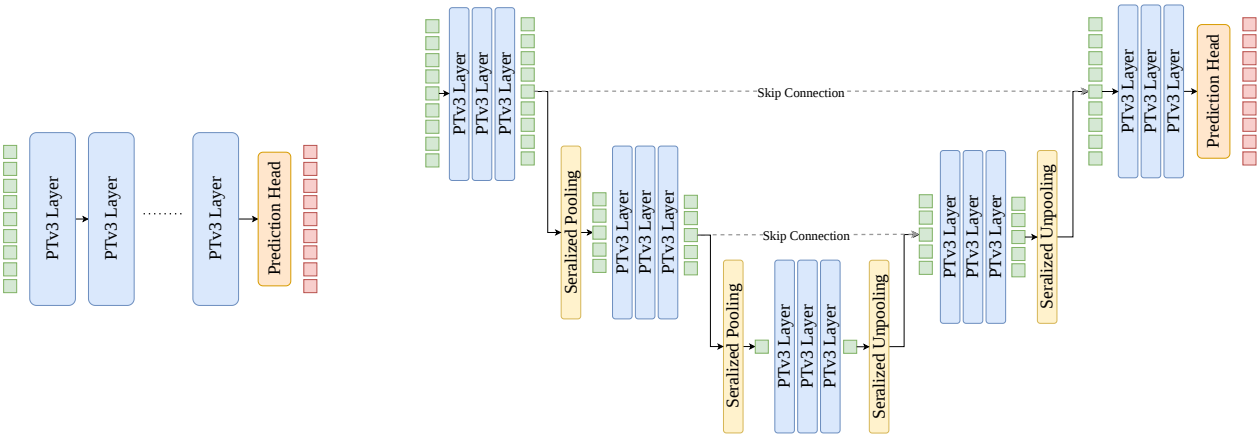
Serialization and Patch-Based Attention A fundamental challenge in applying transformers to 3D point clouds is the unstructured, permutation-invariant nature of the data. Computing global self-attention across 10,000 unorganized points yields an intractable quadratic complexity, $O(N_p^2)$. We leverage the core innovation of PTv3 to resolve this: space-filling curves. Specifically, the 3D coordinates are mapped to a 1D sequence using a z-order (Morton) curve, which preserves spatial locality. The serialized 1D sequence is then divided into local patches, allowing the network to compute highly efficient, localized self-attention rather than expensive global or KNN attention.

Isotropic Architectural Adaptation The architectural backbone of a point-based diffusion model dictates how spatial features are extracted and refined. To optimize generation quality and training stability, we compared two primary design paradigms: the hierarchical U-Net and the isotropic transformer.

The standard PTv3 architecture typically employs a hierarchical U-Net design. As illustrated in Figure 9(b), this approach utilizes serialized spatial pooling (downsampling) and unpooling (upsampling) layers. Consequently, the number of point tokens (N_p) and the feature dimension dynamically change across the network to aggregate multi-scale features from local to global contexts. However, empirical evaluations during the development of GeoPriorPC revealed severe gradient explosion and training instability when applying this hierarchical structure to our specific conditional diffusion setup. The aggressive downsampling inherently discards fine-grained spatial information, which severely complicates the precise coordinate refinement required during the diffusion's reverse process.

Consequently, in this project, we chose the isotropic architecture, aligning with the layout of DiPT. Unlike the hierarchical approach, our isotropic backbone (shown in Figure 9(a)) maintains a strictly constant point resolution ($N_p = 10,000$) and a uniform feature dimension across all PTv3 transformer layers. The

network consists of 6 identical blocks without any intermediate pooling operations. This isotropic design completely mitigated the gradient instability observed in the U-Net variant, seamlessly retaining the powerful localized attention mechanisms necessary for high-fidelity point cloud generation. A comprehensive evaluation and comparison of the results derived from these two architectural designs will be discussed in Section 4.5.4.



a) Isotropic Design b) UNet Design

Figure 9: A diagram explaining the structural differences between the Isotropic and U-Net designs. a) Represents the isotropic design, where we utilize N sequential PTv3 layers. Both the feature dimension (*dim*) and the total number of point tokens (N_p) remain constant throughout the network. b) Represents the U-Net design, where the number of tokens (N_p) and the feature dimension dynamically change across layers through serialized spatial pooling and unpooling operations.

3.5.3 Dual-Conditioning Mechanism for Geometric Priors

The most critical contribution of the GeoPriorPC Stage 2 architecture is how the geometric prior from Stage 1 ($\hat{z}_0$) is injected into the DiPT backbone. Standard point diffusion models condition solely on time and global class labels. To enforce strict structural adherence to the predicted oblique normals, we engineered a novel dual-conditioning mechanism consisting of both global modulation and local cross-attention.

Global Modulation via adaLN-Zero To inform the entire point cloud of the overall building structure, the network first relies on global modulation. In the standard single-view setting, the single dense latent normal map $\hat{z}_0$ is passed through a MLP mapping network, compressing it into a 1D global context vector, F_{global}. This global context is fused with the standard diffusion timestep embedding t and injected into every DiPT block via adaLN-Zero. Rather than learning standard LayerNorm scale and shift parameters, the network regresses them directly from $F_{global+t}$. Crucially, following recent advancements in diffusion transformers, we initialize the final linear layer of this modulation mechanism to zero. This ensures that at the start of training, the DiPT blocks behave as identity functions, preventing the complex global geometry features from destabilizing the early gradient updates.

To optionally support multi-view conditioning, this global pathway is expanded. Instead of a single map, the framework processes four distinct latent normal maps $\{\hat{z}_0^{(v)}\}_{v=1}^{4}$. Each map is independently

compressed by the MLP into a view-specific context vector. These four discrete vectors are then concatenated together to form a comprehensive, multi-view F_{global} that dictates the volumetric shape from all four perspectives simultaneously.

Local Cross-Attention to Latent Tokens While global modulation guides the macro-structure, synthesizing accurate local details (e.g., specific roof slopes and facade boundaries) requires fine-grained spatial alignment. To achieve this, we introduce a local cross-attention layer within each DiPT block.

In the baseline single-view architecture, the latent normal map $\hat{z}_0 \in \mathbb{R}^{h \times w \times c}$ is flattened spatially into a sequence of N 2D tokens (e.g., $N = 1000$). These tokens are augmented with 2D cosine positional embeddings to retain their spatial layout. In the cross-attention mechanism, the 3D point features act as the *Queries* (Q), while these flattened 2D normal tokens act as the *Keys* (K) and *Values* (V).

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

By actively querying the latent normal map at every layer, the 3D points dynamically pull geometric information specific to their localized region.

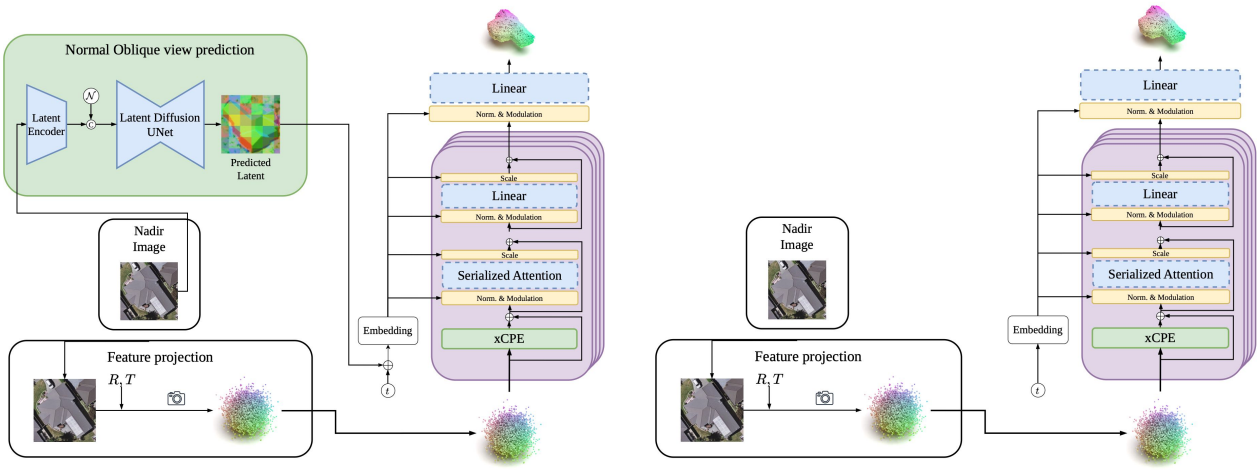
To support the multi-view option, the cross-attention mechanism is scaled to aggregate features across all four perspectives. First, local feature maps are extracted for each of the four latent normal views. To ensure the transformer can explicitly distinguish between the different projection angles, we introduce four learnable view-embedding tokens. Each image's feature map is additively conditioned with its corresponding view token *before* being flattened. Once flattened, each sequence receives its standard 2D cosine positional embedding to preserve intra-view spatial layouts. Finally, the four independent sequences are concatenated along the sequence dimension.

Consequently, the multi-view cross-attention mechanism processes an expanded sequence of $4N$ tokens (e.g., 4000 tokens). This expanded key-value space allows the 3D points to dynamically query geometric information from the most advantageous projection angle for their localized region, without fundamentally altering the underlying attention mathematics.

3.5.4 Ablation Model Variations

To systematically evaluate the individual contributions of the proposed dual-conditioning mechanisms and isolate the impact of the extracted geometric priors, three distinct architectural variations of the framework were engineered for comparative analysis:

GeoPriorPC-adaLN (Global Modulation Only) This variation serves to isolate the efficacy of fine-grained spatial cross-attention (see figure.10a). In this setup, the local cross-attention layers within each DiPT block are entirely disabled. The framework retains the Stage 1 latent oblique normal prior $\hat{z}_0$, but compresses it exclusively into the 1D global context vector F_{global} via the MLP mapping network. This global context is then concatenated with the standard diffusion timestep embedding and injected strictly through the adaLN-Zero layers. Consequently, the model is informed of the macro-structural proportions and building scale, but lacks a mechanism for pixel-to-point local feature matching.



(a) GeoPriorPC-adaLN Architecture

(b) GeoPriorPC-Ignore Architecture

Figure 10: Comparison of the ablation model designs: (a) GeoPriorPC-anaLN handles the latent normal map prior solely as a compressed global context vector via adaLN-Zero layers without local cross-attention, while (b) GeoPriorPC-Ignore completely bypasses the geometric prior stage and relies strictly on the raw 2D input maps.

GeoPriorPC-Ignore (No Geometric Prior) To establish a direct point of comparison showing the baseline performance of our updated transformer backbone without external geometric assistance, the *GeoPriorPC-Ignore* variant completely strips away the Stage 1 intermediate normal map condition (see figure.10b). This model relies strictly on the traditional 2D multi-modal spatial context (the raw nadir RGB image I, the building footprint mask M_b, and the Sobel edge map E). The 3D coordinates are back-projected onto these raw 2D input maps via bilinear interpolation, mirroring the conditioning pipeline of single-stage architectures without utilizing any extracted geometric priors.

Baseline-normal (Time-Embedding Modulation) Engineered purely as an experimental control, this variation investigates an alternative global fusion pathway reminiscent of class-conditional diffusion training layouts. Rather than routing the global geometric context through specialized modulation layers across every transformer block, the 1D global context vector derived from the latent normal map is additively combined directly with the initial time embedding vector at the network's input stage. This setup tests whether injecting the macro-geometric prior as a singular, generalized time-dependent prompt is sufficient for stable coordinate synthesis compared to our finalized dual-conditioning framework.

4 Experiments and Results

This chapter presents a comprehensive empirical evaluation of the GeoPrior-Diff framework. We begin by detailing the experimental setup, including the dataset, evaluation metrics, and the baseline methodologies used for comparison. Following this, we present the primary quantitative results, demonstrating the SOTA performance achieved by integrating a decoupled geometric prior. The chapter then transitions into a qualitative visual analysis to highlight the structural fidelity of the generated 3D building models. Finally, we provide extensive ablation studies and architectural explorations, including multi-view conditioning and aspect ratio analyses, to empirically justify the design decisions formulated in Chapter 3.

4.1 Experimental Setup

4.1.1 Dataset and Preprocessing

The proposed GeoPrior-Diff framework is trained and evaluated utilizing the AIM2PC dataset [1]. AIM2PC provides a rigorous benchmark specifically designed for urban geometric reconstruction, containing paired high-resolution nadir aerial imagery, corresponding building footprint masks, and ground-truth 3D building point clouds. Lacking a canonical benchmark split for its 2,566 total samples, we partitioned the dataset into 2,248 samples for training and 318 for validation. For fair comparison, all baseline models evaluated in this study were retrained from scratch using this exact data split. Each ground-truth building point cloud comprises $N_p = 10,000$ points, which are globally normalized into a unit sphere to maintain numerical stability across the diffusion backbone.

4.1.2 Implementation and Training Details

To ensure reproducibility and geometric stability, the framework employs specific data handling and optimization strategies across both stages.

Data Resolutions and Feature Extraction During Stage 1 (Geometric Prior Training), the input nadir images and corresponding normal maps are resized to 256×256 pixels to fine-tune the LDM. During Stage 2 (3D Generation), the 2D input tensors are resized to 244×244. The local prior encoder Φ_{local} processes the compressed normal map to output an intermediate spatial resolution of $64 \times 64 \times d$, which is subsequently pooled to a global context dimension of $d = 256$.

Zero-Mean Coordinate Centering (CDPM) Standard 3D diffusion models are susceptible to coordinate drift, in which the point cloud's center of mass shifts uncontrollably during the iterative reverse process. To prevent this, we strictly adopt the CDPM [83] recipe. This enforces a zero-mean point cloud distribution across both the forward noise injection and reverse denoising processes, keeping the geometry anchored to the origin.

Optimization and Inference The Stage 2 3D generation models are trained for 100,000 steps using a batch size of 12 on a single NVIDIA L40S GPU. We utilize the AdamW optimizer with a maximum learning rate

of 1×10^{-3}. To manage gradient updates, the baseline models employ a Cosine Annealing scheduler with 2,000 warmup steps, whereas the DiPT-based architectures utilize the OneCycleLR scheduler to maximize convergence stability.

During test-time inference, we utilize DDIM to accelerate sampling. Stage 1 generates the latent normal map $\hat{z}_0$ in $T_1 = 50$ steps using a linear noise schedule ($\eta = 0$) and no classifier-free guidance. Stage 2 then synthesizes the final 3D point cloud in $T_2 = 1000$ DDIM steps. To ensure deterministic evaluation and fair benchmarking, all reported single-sample generations utilize a fixed random seed of 60.

4.1.3 Evaluation Metrics

To quantitatively assess the geometric fidelity of the 3D reconstructions and the accuracy of the 2D geometric priors, we employ the following rigorous mathematical metrics.

3D Geometric Metrics We evaluate the 3D point clouds in their normalized coordinate space using three unitless metrics:

- **Earth Mover's Distance (EMD):** EMD frames the evaluation as an optimal transport problem, measuring the minimum cost required to transform the predicted point cloud $\hat{P}$ into the ground truth P. It heavily penalizes global structural anomalies and mass displacement. For two point sets of equal size, it is defined by finding a bijection $\phi : P \to \hat{P}$:

$$d_{EMD}(P,\hat{P}) = \min_{\phi:P\to\hat{P}} \frac{1}{|P|} \sum_{p\in P} ||p - \phi(p)||_2 \tag{9}$$

- **Density-Aware Chamfer Distance (DCD):** While standard Chamfer Distance computes the nearest-neighbor Euclidean distance between point sets, it is blind to density variations and can be easily skewed by localized clustering. DCD penalizes uneven point distributions, ensuring that the predicted points uniformly cover the architectural surfaces without artifacting.

- **F1-Score (F1@0.02):** To balance spatial precision and recall, we compute the F1-score. A predicted point is considered a "true positive" if it falls within a strict distance threshold ($\tau = 0.02$) of a ground-truth point:

$$F1@0.02 = \frac{2 \cdot \text{Precision}(\tau) \cdot \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)} \tag{10}$$

2D Normal Map Metrics To evaluate the quality of the oblique geometric priors predicted by the fine-tuned LDM in Stage 1, we calculate the pixel-wise angular deviation between the predicted normal vectors $\hat{n}_i$ and the ground-truth normal vectors n_i.

- **Mean Absolute Error (MAE):** Reports the average angular error across all valid building pixels, measured in degrees:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \arccos(n_i \cdot \hat{n}_i) \times \frac{180}{\pi} \tag{11}$$

- **Sub 11.25° Accuracy:** Measures the strict percentage of pixels where the angular error is bounded strictly below 11.25°, indicating high-fidelity geometric alignment.

4.2 Main Quantitative Results

To rigorously evaluate the efficacy of the GeoPriorPC pipeline, we benchmark our results against several leading single-image 3D reconstruction paradigms. These include purely convolutional or transformer-based direct prediction networks, as well as SOTA single-stage diffusion baselines (specifically PC2 [87], CCD-3DR [83], and AIM2PC [1]) that attempt to map nadir RGB features directly to 3D coordinates without an explicit intermediate geometric prior.

4.2.1 Global Metric Comparison

The quantitative performance of GeoPriorPC against the established baselines is summarized in Table 3. By decoupling the generation process and explicitly conditioning the 3D diffusion backbone on the latent oblique normal maps, our proposed framework achieves a significant leap in SOTA performance across all primary geometric metrics.

Table 3: Quantitative comparison of single-image 3D building reconstruction on the AIM2PC dataset [1]. Arrows indicate whether higher (↑) or lower (↓) scores are optimal. Best results are highlighted in bold.

Model	**F1@0.02** (↑)	**EMD** (↓)	**DCD** (↓)
PC2 [87]	0.484	0.00310	0.730
CCD-3DR [83]	0.512	0.00280	0.711
AIM2PC [1]	0.524	0.00255	0.696
GeoPriorPC (Ours)	**0.754**	**0.00131**	**0.598**

4.2.2 Performance Analysis

As evidenced by the global metrics, single-stage baselines consistently struggle to resolve the inherent vertical ambiguity of the top-down view. Because these models rely exclusively on explicit 2D spatial features (RGB and edge maps), they lack the 3D volumetric context required to infer out-of-plane structures. Mathematically, this manifests in high EMD scores, indicating that the predicted mass is incorrectly distributed, often resulting in flattened building volumes or collapsed facades.

GeoPriorPC fundamentally resolves this bottleneck. By providing the DiPT backbone with a reliable structural anchor (the $\hat{z}_0$ normal prior), the generative burden is significantly reduced. This explicit geometric guidance yields a massive 44% relative improvement in the overall F1-score compared to the previous State-of-the-Art (AIM2PC).

Furthermore, the structural consistency enforced by our dual-conditioning mechanism results in a remarkable 49% reduction in EMD. This dramatic drop indicates that the model not only captures the

macroscopic layout of the building footprints but also accurately recreates complex, fine-grained local densities such as perpendicular wall drops and sharp roof pitches. The corresponding reduction in DCD from 0.696 to 0.598 confirms that our generated surfaces are uniform and artifact-free.

4.3 Qualitative Evaluation

While quantitative metrics provide a global assessment of geometric distribution, visual qualitative analysis is crucial for evaluating architectural realism and structural fidelity.

4.3.1 Visual Comparisons and Structural Fidelity

To fully contextualize the performance of the framework, it is necessary to visually evaluate both the intermediate geometric prior (Stage 1) and the final 3D point cloud generation (Stage 2).

As illustrated in Figure 11, the fine-tuned LDM successfully generates the out-of-plane geometry from the top-down aerial perspective. Despite only receiving a flat nadir RGB image as input, the predicted oblique normal maps accurately capture complex multi-plane roof structures. The distinct color variations within the generated normals clearly define sharp planar intersections, varying roof slope inclinations, and, most crucially, the vertical drop of the building facades. This visual evidence confirms that the LDM successfully learned the underlying 3D architectural distributions rather than simply copying 2D image textures.

The impact of this high-fidelity geometric prior becomes visually evident when examining the final 3D coordinate generation (Figure 12). Models operating without a geometric prior, such as AIM2PC, successfully identify the general 2D building footprint but frequently fail to extrude the geometry correctly. These baseline predictions exhibit severe ”pancaking” effects, where the Z-axis variance is improperly compressed, and multi-level roof structures are blurred into a single, noisy continuous surface.

Conversely, GeoPriorPC demonstrates a high degree of structural fidelity. By dynamically ”attending” to the distinct planes defined in the predicted oblique normal map during the denoising process, the network accurately synthesizes distinct architectural features. Our predictions exhibit sharp, well-defined roof ridges, precise slope angles, and sheer vertical facades that closely mirror the ground-truth topologies, entirely resolving the flattening artifacts seen in the baseline methods.

4.4 Complexity and Inference Efficiency

Given the transition from a traditional 3D convolutional backbone (PVCNN) to a transformer-based backbone (DiPT), evaluating the computational overhead is essential. All timing evaluations were conducted on a single NVIDIA L40S GPU. For the transformer models, we maintained a consistent architecture depth of 6 blocks and a hidden dimension of $d = 256$.

As detailed in Table 4, the full GeoPriorPC model contains 27.56M parameters and achieves an average iteration time of 0.154 seconds. Despite the self-attention mechanisms inherent to transformers, our backbone remains computationally lighter and faster than the PVCNN-based AIM2PC baseline (28.05M parameters, 0.160 s/iter). This efficiency is directly attributable to the space-filling curve serialization utilized

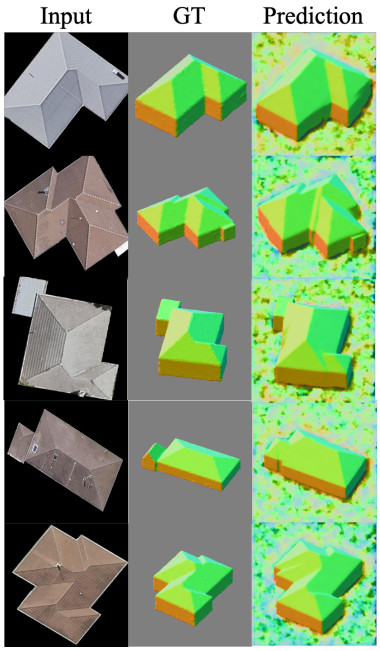


Figure 11: Qualitative results of finetuned LDM. It shows that the model can successfully generate the normal oblique view from just a single nadir image.

by the PTv3 blocks, which allows for localized patch-based attention rather than computationally prohibitive global attention computations.

Table 4: Comparison of model complexity and average iteration time evaluated on an NVIDIA L40S GPU.

Model Variation	**Parameters (M)**	**Time (s/iter)**
AIM2PC Baseline	28.05	0.160
GeoPriorPC-Concat	29.73	0.122
GeoPriorPC (Ours)	27.56	0.154

4.5 Ablation Studies and Architectural Explorations

To empirically justify the design decisions formulated in Chapter 3, we conducted extensive ablation studies. These experiments isolate the impact of the pre-trained weights, the dual-conditioning mechanism, and several architectural variants explored during development.

4.5.1 Impact of Pre-trained Latent Priors (Stage 1)

To validate the necessity of leveraging foundational models for the geometric prior, we compared our fine-tuned Marigold LDM against an identical architecture initialized from scratch with random weights.

As shown in Table 5, the model trained from scratch completely fails to learn the complex mapping from nadir textures to out-of-plane normals, resulting in a severe MAE of 49.5°. Visually, these predictions collapse into meaningless noise (see figure 13). In contrast, fine-tuning the pre-trained Stable Diffusion



Figure 12: Qualitative comparison of 3D building reconstructions. From left to right: Input nadir image, Ground Truth, Our model with global context only, our full GeoPriorPC model, and AIM2PC baseline.

weights allows the network to leverage its generalized world knowledge, achieving an MAE of 19.3° and capturing 58.3% of pixels within a strict 11.25° error threshold. This confirms that large-scale pre-training is indispensable for resolving the structural ambiguities of aerial imagery.

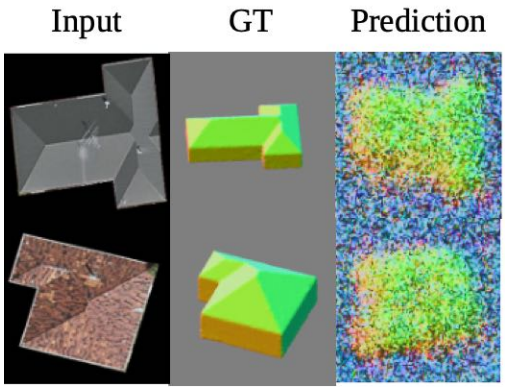


Figure 13: Results from an LDM trained from scratch. The results show that the model completely collapsed and couldn't generate any meaningful output.

Table 5: Evaluation of Stage 1 normal map prediction quality, comparing random initialization versus fine-tuning pre-trained LDM weights.

Model Initialization	**Sub 11.25° (%) (↑)**	**MAE (deg) (↓)**
Baseline (From Scratch)	5.5	49.5
Fine-tuned LDM (Ours)	58.3	19.3

4.5.2 Backbone and Conditioning Strategies (Stage 2)

We isolated the contributions of the DiPT backbone and the dual-conditioning mechanism as discussed in section 3.5.4. As shown in Table 6, simply upgrading the AIM2PC baseline to the DiPT architecture without any normal conditioning (GeoPriorPC-Ignore) improved the F1-score from 0.523 to 0.565, proving the transformer processes unstructured point features more effectively than 3D convolutions.

Injecting the geometric prior strictly via a global context vector (GeoPriorPC-Concat) improved the score to 0.678. However, global context alone lacks fine-grained spatial alignment. The full model, which utilizes both global adaLN-Zero injection and local cross-attention to match 3D points to 2D normal tokens, achieved the highest F1-score of 0.754. This confirms that active local feature matching is necessary to sculpt accurate architectural details.

4.5.3 Architectural Exploration: Single-View vs. Multi-View Conditioning

During development, we hypothesized that providing Stage 2 with multiple geometric priors from different perspectives would yield a more comprehensive structural understanding of the building. To test this, we generated four distinct normal views by varying the rendering yaw angle $\{0°, 90°, 180°, 270°\}$ and trained the LDM to predict this composite condition, more details explained in section 3.4.4.

Table 6: Ablation study on architectural backbones and prior conditioning strategies.

Model Variation	**F1@0.02 (↑)**	**EMD (↓)**	**DCD (↓)**
Baseline Backbone (PVCNN)			
Baseline (No Prior)	0.523	0.00255	0.696
Baseline-normal	0.560	0.00220	0.682
Proposed Backbone (PTv3)			
GeoPriorPC-Ignore (No Prior)	0.565	0.00216	0.683
GeoPriorPC-adaLN (Global Only)	0.678	0.00187	0.633
GeoPriorPC (Global + Local)	**0.754**	**0.00131**	**0.598**

A purely global metric evaluation initially suggested that the single-view approach (0°) was superior, achieving an overall F1-score of 0.739 compared to 0.725 for the multi-view variant (Table 7). As detailed extensively in Chapter 5, this performance drop is attributed to cross-view geometric inconsistency, where the LDM generates conflicting spatial cues across the 2×2 grid, destabilizing the DiPT cross-attention mechanism.

However, a granular, per-sample analysis revealed a highly nuanced relationship between the two approaches. As visualized in the scatter plot in Figure 14, we mapped the F1 scores of the single-view and multi-view models on a per-sample basis. The statistical correlation between the two models is notably weak regarding structural fidelity (F1 Spearman $\rho = 0.376$), indicating that the two models succeed and fail on entirely different distributions of architectural topologies.

Table 7: Quantitative comparison of Single-View and Multi-View conditioning. The "Oracle" column represents the theoretical maximum if the best prediction per sample is dynamically selected, proving the complementary nature of the two methods.

Metric	**Single-View (S)**	**Multi-View (M)**	**Oracle (S ∪ M)**	**Spearman ρ**
F1@0.02 (↑)	0.739	0.725	**0.805**	0.376
EMD (↓)	0.0010	0.0013	**0.0009**	0.662
DCD (↓)	0.612	0.589	**0.565**	0.437

To quantify this complementary behavior, we constructed an "Oracle" model representing the theoretical upper bound if the framework could perfectly select the better prediction (Single or Multi) for each specific test sample. As shown in Table 7, the Oracle framework yields a massive performance leap, jumping to an F1-score of 0.805. Furthermore, while the multi-view approach lags in F1, it intrinsically achieves a better Density-Aware Chamfer Distance (0.589 vs 0.612), suggesting it distributes points more uniformly across complex facades even if global alignment occasionally drifts.

This complementary dynamic is visually corroborated in Figure 15, which projects the prediction error as a color map directly onto the generated 3D point clouds. In the first example, the multi-view model vastly outperforms the single-view baseline, successfully utilizing the supplementary angles to resolve complex undercuts and sheer vertical facades that are completely occluded from the 0° perspective. Conversely,

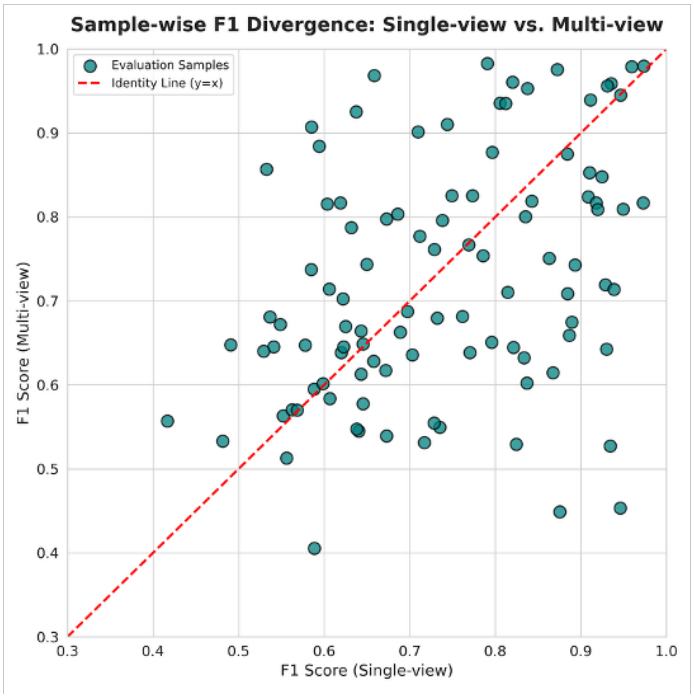


Figure 14: A scatter plot showing F1-score for both single and multiview predictions. Each point represents a test sample, while the x-axis represents the f1-score for the single-view model and the y-axis represents the f1-score for the multiview model.

in the second example, the single-view model excels, maintaining crisp, stable roof ridges where the multi-view prediction suffers from structural noise due to conflicting prior signals.

Ultimately, these findings demonstrate that while a single, high-fidelity normal prior provides the most stable global baseline, integrating multi-view geometries offers a massive, untapped ceiling for generative accuracy. Future work should focus on engineering a ”cross-view consistency loss” during the Stage 1 LDM generation to merge these complementary strengths without destabilizing the Stage 2 attention mechanisms.

4.5.4 Architectural Exploration: Isotropic vs. Hierarchical DiPT Backbone

We also investigated a U-Net style hierarchical architecture for the DiPT backbone, mirroring the original layout of PTv3. This design incorporates aggressive spatial downsampling and upsampling layers to aggregate multi-scale features, contrasting with the constant-resolution approach of our finalized isotropic layout.

Empirical observations revealed that this hierarchical variant was highly susceptible to severe gradient explosion during the reverse diffusion process. While point cloud diffusion is inherently sensitive, the gradient flow in the isotropic baseline (which maintains a constant $N_p = 10,000$ resolution across all blocks) could be strictly contained via precise batch sizing and learning rate scheduling (e.g., OneCycleLR). Conversely, as illustrated in Figure 16a, the U-Net variant consistently suffered from extreme gradient volatility across multiple hyperparameter configurations. While the isotropic model maintains a bounded, stable clipped gradient norm throughout training, the U-Net design exhibits continuous, violent spikes that ulti-

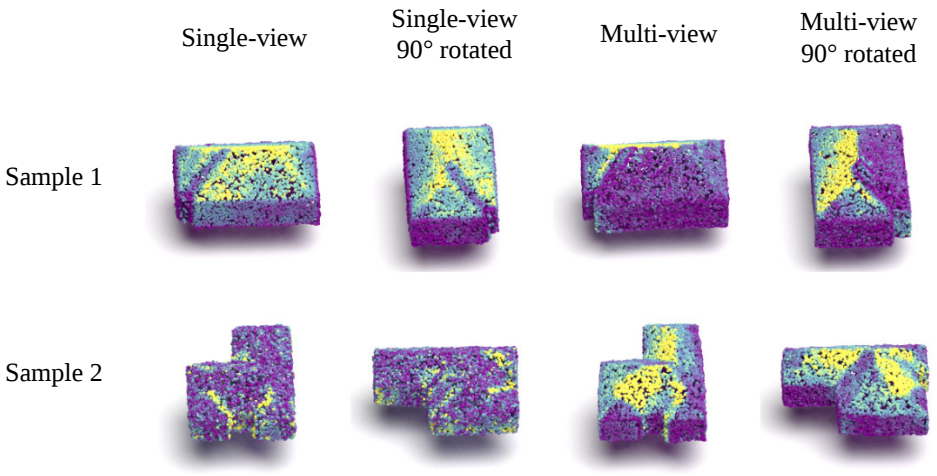


Figure 15: A scatter plot showing F1-score for both single and multiview predictions. Each point represents a test sample, while the x-axis represents the f1-score for the single-view model and the y-axis represents the f1-score for the multiview model.

mately derail the optimization process.

This instability directly impacts the model's convergence and generative capabilities, as demonstrated by the Validation Earth Mover's Distance (EMD) in Figure 16b. Interestingly, the U-Net architecture demonstrates a highly promising initial learning trajectory, achieving a sharp and competitive reduction in EMD during the very early stages of training. However, coinciding with the onset of the severe gradient spikes, the U-Net completely loses its capacity to further refine the geometric coordinates, plateauing at a sub-optimal state or diverging. In stark contrast, the stable gradient dynamics of the isotropic design allow it to smoothly and continuously minimize the EMD.

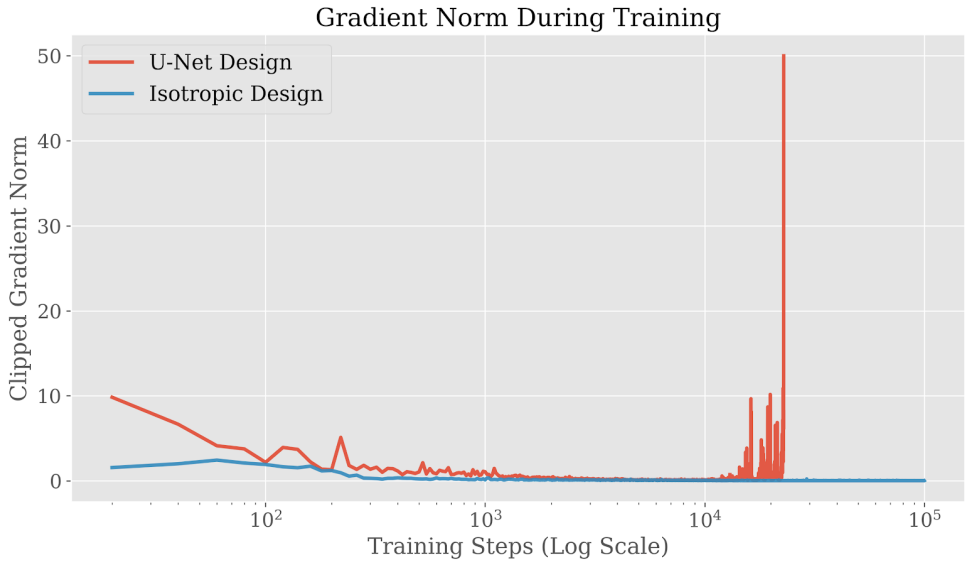
Given the robust, stable performance of the isotropic design and the constraints of the current research scope, the 6-block isotropic layout was finalized as the core generative backbone for this work. Nevertheless, the strong initial convergence of the hierarchical model before the onset of instability suggests that the U-Net framework holds significant potential. Thoroughly diagnosing the root causes of this spatial pooling-induced instability and successfully adapting the U-Net architecture to handle these diffusion dynamics remains a highly promising avenue for future work, as it could further enhance multi-scale feature representation and computational efficiency.

4.5.5 Architectural Exploration: Explicit Aspect Ratio Conditioning

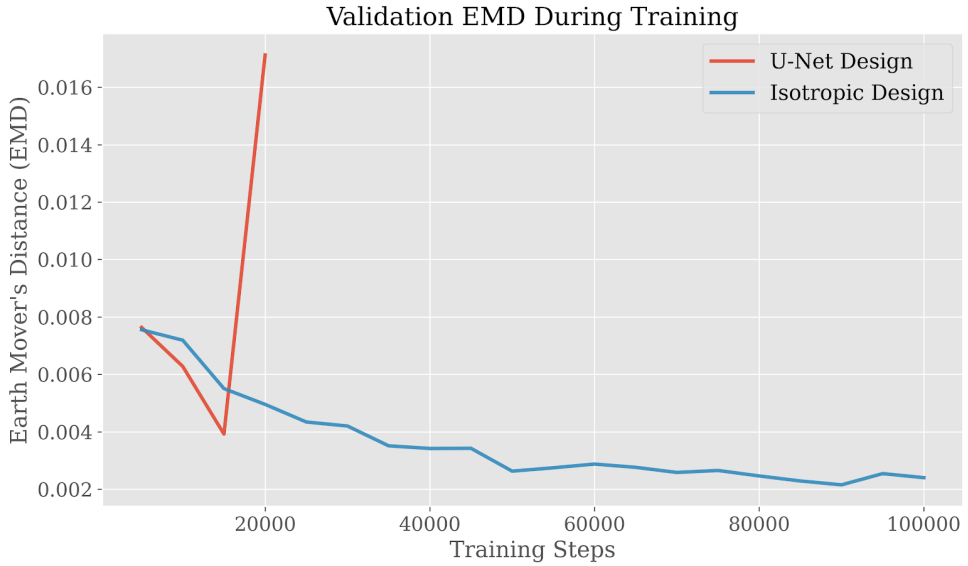
To address varying image proportions, we tested an explicit aspect ratio conditioning module. A two-layer MLP projected the input aspect ratio into a high-dimensional vector, which was additively combined with the time embedding.

As shown in Table 8, while this module provided a marginal improvement in Stage 1 normal prediction, it yielded negligible improvements in the final Stage 2 3D generation.

This redundancy is attributed to our two-stage training curriculum. By initially training the network on forcibly resized 256×256 images and subsequently fine-tuning it on unconstrained aspect ratios, the



(a) Clipped Gradient Norm (Log Scale)



(b) Validation Earth Mover's Distance (EMD)

Figure 16: A comparison of training dynamics between the hierarchical U-Net and the Isotropic designs. (a) Displays the clipped gradient norm across training steps, highlighting the severe gradient explosion and volatility inherent to the U-Net design compared to the bounded stability of the isotropic model. (b) Illustrates the resulting impact on model convergence. While the U-Net shows a strong initial decrease in error, the subsequent gradient instability prevents it from effectively refining the geometric coordinates, whereas the isotropic architecture smoothly minimizes the Earth Mover's Distance.

Table 8: Impact of explicit aspect ratio embedding on Stage 1 and Stage 2 performance.

	Stage 1	Stage 2		
Configuration	**MAE** (↓)	**F1@0.02** (↑)	**EMD** (↓)	**DCD** (↓)
Implicit Curriculum	19.3	0.754	0.00131	0.598
Explicit MLP Embedding	18.7	0.755	0.00130	0.595

LDM implicitly learned to process varying spatial proportions seamlessly. Therefore, the explicit MLP module was discarded to reduce architectural complexity.

4.6 Architectural Exploration: Effect of Random Seeds in the Multi-View Setting

To ensure the replicability of our proposed GeoPriorPC framework and to investigate the inherent stochasticity of the diffusion process, we conduct a series of experiments evaluating the impact of random seeds. Specifically, we utilize three distinct seeds, **43, 100, and 2039**, across the multi-view setting. It is important to note that all models utilized in this section are already fully trained, and the evaluations are performed strictly in inference mode.

4.6.1 Stochasticity in Stage 2 (Image-to-Point Generation)

First, we isolate the stochasticity of the Stage 2 DiPT backbone. In this experiment, we run Stage 2 on the exact same input (using a fixed Stage 1 geometric prior) multiple times using the three different random seeds.

As shown in Table 9, the performance of the model remains virtually identical across all three runs. To quantify the maximum potential of this variation, we computed an "Oracle" result, which dynamically selects the best prediction per sample based strictly on the F1-score. The Oracle metrics show only a negligible improvement. This indicates that the 3D denoising process in Stage 2 is highly stable; merely varying the random noise during the point cloud generation phase does not yield any meaningful structural diversity or performance gain.

Table 9: Effect of Random Seeds on Stage 2 Inference

Configuration	**F1-score** (↑)	**EMD** (↓)	**DCD** (↓)
Run 1 (Seed 43)	0.724	0.00130	0.589
Run 2 (Seed 100)	0.730	0.00125	0.587
Run 3 (Seed 2039)	0.725	0.00129	0.587
Oracle (Selected via F1)	**0.734**	**0.00122**	**0.585**

4.6.2 Stochasticity in Stage 1 (Latent Geometric Prior Extraction)

Next, we evaluate the framework's sensitivity to random seeds during the Stage 1 LDM inference. In this setup, we run Stage 1 with the three random seeds to generate three distinct normal priors, which are then

individually passed to Stage 2.

While the performance of any individual run remains relatively consistent, the Oracle results demonstrate a significant shift. As shown in Table 10, calculating the Oracle performance across the three runs yields a substantial increase in the F1-score (jumping to 0.781) and corresponding reductions in EMD and DCD. This reveals that the variability in the Stage 1 LDM predictions provides highly complementary geometric cues. The generative diversity of the foundational prior introduces a much higher performance upper bound for the overall framework based on how many times Stage 1 is run.

Table 10: Effect of Random Seeds on Stage 1 Prior Generation

Configuration	**F1-score (↑)**	**EMD (↓)**	**DCD (↓)**
Run 1 (Seed 43)	0.724	0.00130	0.589
Run 2 (Seed 100)	0.740	0.00113	0.580
Run 3 (Seed 2039)	0.732	0.00111	0.582
Oracle (3 Runs)	**0.781**	**0.00100**	**0.567**

4.6.3 Addressing the Upper Bound: Test-Time Ensembling

Given the massive performance potential revealed by the Stage 1 Oracle, we explored methods to practically harness this generative diversity. Because the stochastic nature of diffusion inference yields varying predictions based on initialization noise, we applied the test-time ensembling method proposed in the Marigold framework [16].

However, as detailed in Table 11, applying this ensembling technique yields an F1-score of 0.737. While this slightly outperforms an average single run, it falls significantly short of the 0.781 Oracle upper bound. As discussed in Chapter 5, standard ensembling effectively washes out necessary high-frequency structural variations. Developing an intelligent routing mechanism to fully harness this generative diversity remains a primary direction for future work.

Table 11: Comparison of Ensembling vs. Oracle Upper Bound

Method	**F1-score (↑)**	**EMD (↓)**	**DCD (↓)**
Marigold Test-Time Ensemble	0.737	0.00116	0.579
Oracle Upper Bound	**0.781**	**0.00100**	**0.567**

5 Limitations

While the GeoPriorPC framework establishes a new SOTA for single-view 3D building reconstruction, the architectural design and reliance on generative diffusion models introduce several methodological and computational constraints.

Cascading Error Propagation in the Dual-Stage Architecture

The primary vulnerability of the GeoPriorPC pipeline stems from its decoupled, two-stage nature. Because the 3D diffusion process in Stage 2 is explicitly anchored to the latent normal map via cross-attention, the final point cloud is entirely dependent on the geometric accuracy of Stage 1. If the LDM hallucinates an incorrect or physically impossible normal map, frequently triggered by heavy tree occlusions over building footprints or highly irregular, non-rectilinear architecture, the DiPT backbone will faithfully reconstruct those topological errors in 3D. Consequently, geometric inconsistencies in the 2D prior systematically dictate structural artifacts in the final volumetric output.

To empirically quantify these failure modes, a manual inspection was conducted on the lowest-performing 10% of the test set, isolated by F1-score (comprising exactly 32 samples). This error stratification revealed that nearly half of the critical failures (15 samples) originated directly from Stage 1, where the LDM produced degraded or structurally compromised normal maps similar to examples in Figure 17. In 9 of the samples, the Stage 1 prior was sufficiently accurate, but the failure occurred during Stage 2, where the DiPT backbone struggled to correctly synthesize the 3D coordinates from the provided features. The remaining 8 samples suffered from systemic geometric deviations, including spatial misalignments or the generation of disproportionately larger walls and taller building volumes compared to the ground truth. This distribution clearly confirms that while the 3D generative process presents its own distinct challenges, the fidelity of the intermediate 2D prior remains the predominant bottleneck in the pipeline. Figure 17 shows visual examples of this cascading error propagation.

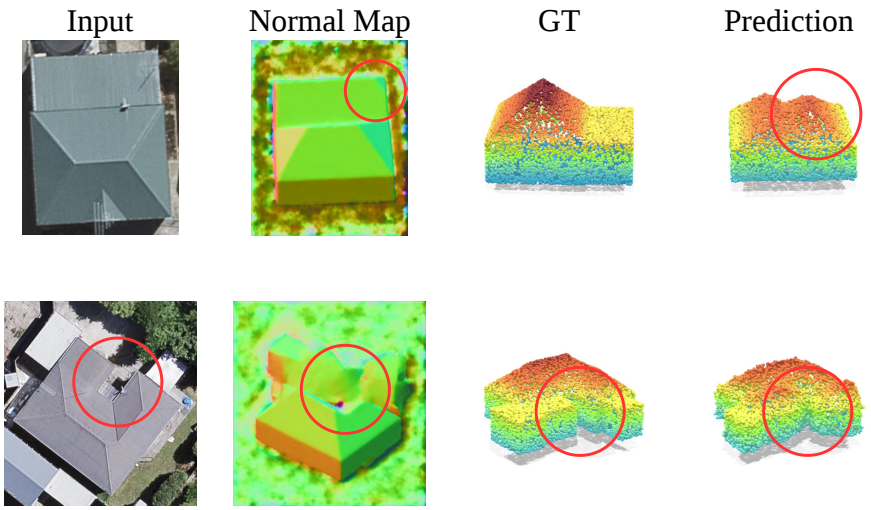


Figure 17: Failure cases demonstrating error propagation. Geometric inaccuracies in the LDM-predicted normal maps dictate the structural artifacts in the final 3D point clouds.

5.1 Cross-View Geometric Inconsistency

Efforts to scale the framework's conditioning mechanism to utilize multiple oblique views simultaneously revealed significant limitations in cross-view coherence. During architectural explorations, the LDM was tasked with predicting four distinct projection angles (0°, 90°, 180°, and 270°) in a single-shot inference step. This was achieved by formatting the output as a 2×2 composite image grid, where each quadrant represented a specific view, to serve as a comprehensive geometric prior. However, the spatial self-attention mechanisms within the LDM's U-Net lacked an explicit mathematical understanding of 3D epipolar geometry. Consequently, despite generating the grid simultaneously, the network predicted the views with conflicting spatial cues. For example, a predicted roof pitch in the 0° quadrant frequently contradicted its corresponding projection in the 90° quadrant. These geometric contradictions inherently destabilized the local cross-attention mechanism of the DiPT during the 3D denoising phase, ultimately yielding lower global F1-scores compared to utilizing a single, high-fidelity view.

5.2 Generative Stochasticity and the "Oracle" Gap

Empirical evaluations isolated a substantial, yet currently untapped, performance ceiling tied to the inherent stochasticity of diffusion inference. While the Stage 2 DiPT backbone demonstrated high stability across random seeds, the Stage 1 LDM exhibited massive structural variation. Generating multiple geometric priors using different random noise initializations yielded highly complementary structural cues. As demonstrated by an "Oracle" evaluation—dynamically selecting the best prediction per sample—this generative diversity pushes the theoretical F1-score upper bound to 0.781. However, current aggregation methods fail to harness this potential. Standard test-time ensembling simply averages the coordinate predictions, which effectively washes out the high-frequency structural variations and plateaus at an F1-score of 0.737, falling significantly short of the framework's true generative capacity.

5.3 Dataset Bias and Inference Overhead

The supervised fine-tuning of this framework relies on the AIM2PC benchmark, which is constrained to 2,566 samples and is heavily biased toward standard residential topologies. This dataset bias restricts the model's ability to reliably generalize to unrepresented urban morphologies, such as complex brutalist overhangs, modern industrial facilities, or curved facades. Furthermore, the operational reliance on Markovian denoising introduces a severe computational bottleneck. The pipeline requires 50 sequential steps for Stage 1 and 1,000 steps for Stage 2. This high inference overhead fundamentally precludes the current iteration of GeoPriorPC from being deployed for latency-sensitive applications, such as real-time global mapping or rapid disaster response.

5.4 The Transitional Nature of Unstructured Representations

Finally, while unstructured point clouds serve as an excellent mathematical scaffold to bypass rigid topological commitments during generative synthesis, they remain a transitional data format. The generated 3D

coordinates lack explicit topological connectivity, semantic parsing, and watertight surface boundaries. Consequently, these raw point clouds cannot be natively ingested by downstream Building Information Modeling (BIM) systems, urban microclimate simulations, or CAD software without extensive subsequent reverse-engineering into highly parameterized Boundary Representation (B-rep) models.

6 Conclusion and Future Work

6.1 Conclusion

The rapid expansion of UDTs and smart city simulations has created an unprecedented demand for highly accurate 3D urban geometry. While nadir-view aerial imagery provides an abundant and globally accessible data source, extracting full 3D structures from this top-down perspective is a fundamentally ill-posed problem. Traditional deterministic models and single-stage generative networks consistently fail to resolve the inherent vertical ambiguity, resulting in flattened geometric representations that destroy complex roof topologies and vertical facades.

To overcome this representation gap, this thesis introduced GeoPriorPC, a novel two-stage generative framework that seamlessly bridges the dimensional divide between 2D aerial pixels and unstructured 3D point clouds. Rather than forcing a single network to blindly hallucinate missing out-of-plane dimensions, GeoPriorPC systematically decouples the reconstruction task.

In the first stage, the framework successfully adapts the generalized "world prior" of foundational LDMs for geospatial applications. By fine-tuning a pre-trained LDM, the pipeline translates flat nadir RGB inputs into highly structured oblique surface normal maps, explicitly resolving the depth ambiguity without incurring the computational overhead of pixel-space decoding. In the second stage, GeoPriorPC introduces a DiPT backbone based on the PTv3 architecture. To physically ground the 3D generation, a novel dual-conditioning mechanism was engineered. By integrating the latent normal map globally via adaLN-Zero and locally through a spatially-aware cross-attention module, the 3D denoising process is explicitly forced to attend to the structural cues of the generated 2D prior.

Extensive empirical evaluations on the AIM2PC benchmark validate the efficacy of this decoupled approach. GeoPriorPC establishes a new SOTA for single-view 3D building reconstruction. Compared to existing baselines, the proposed framework achieves a 44% relative improvement in the F1-score and reduces the Earth Mover's Distance (EMD) by nearly 49%.

Crucially, this substantial reduction in EMD translates to highly realistic structural fidelity. By successfully avoiding the "pancaking" artifacts prevalent in earlier models, GeoPriorPC accurately synthesizes sharp roof ridges, precise slope angles, and sheer vertical facades. This structural integrity is paramount for downstream UDT applications. Highly precise vertical morphology directly dictates the reliability of CFD simulations for urban microclimates, pedestrian wind comfort, and disaster resilience models, such as storm surge routing. Ultimately, GeoPriorPC proves that injecting high-fidelity geometric priors from foundational 2D image models into highly scalable 3D sequence transformers provides a robust, scalable pathway for extracting these critical 3D manifolds globally, bypassing the reliance on expensive active LiDAR acquisitions.

6.2 Future Work

While GeoPriorPC significantly advances the capability of single-view 3D urban reconstruction and establishes a robust baseline for extracting out-of-plane geometries, the architectural explorations and empirical limitations identified in this research highlight several highly complex, unresolved challenges. Addressing these bottlenecks will require advancing both the generative priors and the fundamental mechanics of 3D

point cloud diffusion. The following sections outline the critical theoretical and methodological avenues for future research.

6.2.1 Scaling Data Diversity and Mitigating Propagation Errors

Because GeoPriorPC is a sequential, dual-stage framework, the final 3D point cloud generation is entirely bottlenecked by the accuracy of the intermediate latent representation. Our error analysis (Section 5) demonstrated that if the Stage 1 LDM hallucinates an incorrect normal map, frequently caused by heavy tree occlusions over building footprints or highly irregular, non-rectilinear architecture, these geometric anomalies propagate directly into the DiPT backbone, resulting in corrupted 3D structures.

To mitigate this, future work must radically scale the training distribution. The AIM2PC dataset, while foundational, is limited to 2,566 samples, heavily biased toward standard residential topologies. Future iterations should explore integrating procedurally generated, synthetic urban datasets (using engines such as CityEngine or Blender) to create an infinitely scalable, highly diverse training corpus featuring curved facades, complex brutalist overhangs, and diverse global architectural styles.

Furthermore, to explicitly handle occlusions, future architectures should integrate multi-modal semantic gating. By incorporating off-the-shelf semantic segmentation models to identify vegetation and shadow masks, the Stage 1 LDM could be modified with an explicit cross-attention penalty. This would dynamically reduce the attention weights over occluded pixels, forcing the network to hallucinate the missing geometry based purely on the unoccluded architectural context rather than attempting to extrude a tree canopy as a building feature.

6.2.2 Harnessing the Generative Oracle: Uncertainty-Aware Routing

Our empirical evaluation of the diffusion process's stochasticity revealed a substantial, untapped generative ceiling. The "Oracle" analysis (Section 4.6) demonstrated that dynamically selecting the optimal geometric prior from multiple Stage 1 random seeds yields an exceptional F1-score of 0.805. However, standard test-time ensembling mathematically failed to reach this bound, as naive averaging destroys the high-frequency structural variations that make a specific seed accurate.

Unlocking this upper bound requires abandoning simple ensembling in favor of intelligent, uncertainty-aware routing. Future research should design a lightweight, auxiliary "Critic Network" positioned between Stage 1 and Stage 2. This network would take N stochastically generated latent normal maps as input and score them based on physical plausibility and boundary coherence. By evaluating the structural entropy or calculating a metric akin to a Fréchet Inception Distance (FID) for surface normals, this routing module could dynamically filter out catastrophic hallucinations and pass only the highest-confidence geometric prior to the computationally heavy 3D denoising stage.

6.2.3 Enforcing Epipolar Consistency in Multi-View Generation

During the architectural explorations detailed in Chapter 4, the multi-view conditioning approach (generating four distinct oblique normal maps at $0^\circ, 90^\circ, 180^\circ$, and 270°) failed to yield global performance gains. In this

approach, the LDM was trained to predict all four views simultaneously by concatenating them into a single 2×2 composite image grid. However, a fundamental flaw remained: cross-view geometric inconsistency. Although the views were generated in a single diffusion pass, the standard convolutions and spatial self-attention mechanisms within the LDM's U-Net lacked an explicit mathematical understanding of the 3D projection relationships between the grid quadrants. Consequently, the resulting normal maps frequently contained conflicting spatial cues, such as a roof pitch in the 0° view failing to align with its corresponding projection in the 90° view, which inherently destabilized the local cross-attention mechanism of the Stage 2 transformer.

To successfully harness multi-view priors, future frameworks must enforce strict 3D consistency within this 2D grid generation phase. Since the views are already generated jointly, future architectures should replace standard spatial attention with a 3D-aware Cross-View Attention Module. This mechanism would explicitly map feature queries in one quadrant to their corresponding epipolar lines in the other quadrants, forcing the network to correlate the views geometrically rather than just texturally. Additionally, incorporating a differentiable epipolar or multi-view consistency loss during the diffusion fine-tuning phase would mathematically penalize predictions where the geometry of one facade contradicts another, ensuring the intermediate 2×2 prior forms a cohesive, watertight mathematical shell before any 3D points are generated.

6.2.4 Stabilizing Hierarchical Transformers for Diffusion Dynamics

We investigated a U-Net style hierarchical architecture for the DiPT backbone, which incorporated aggressive spatial downsampling and upsampling layers to aggregate multi-scale features. While this architecture exhibited a highly promising initial learning trajectory, dropping the EMD rapidly in early epochs, it ultimately collapsed due to severe gradient explosion (Figure 16a).

Investigating and resolving the root cause of this instability is critical for scaling 3D point cloud generation. The volatility is likely caused by the interaction between the serialization space-filling curves and the injected Gaussian noise; pooling heavily noised coordinates severely distorts the local variance, causing massive activation spikes. Future work should explore integrating modern vision-transformer stabilization techniques, such as LayerScale and DropPath (stochastic depth), directly into the DiPT blocks. Furthermore, researchers must design a diffusion-specific spatial pooling operation that strictly adheres to the CDPM principles, ensuring that the center-of-mass remains perfectly anchored at zero even as the point resolution is aggressively downsampled.

6.2.5 Inference Acceleration via Diffusion Distillation

While the PTv3 backbone is highly memory-efficient, operating with linear complexity, the fundamental iterative nature of the Markovian denoising process remains a severe computational bottleneck. Requiring 50 steps for Stage 1 and 1000 steps for Stage 2 precludes the GeoPriorPC framework from being deployed for rapid disaster response or city-scale real-time mapping.

A primary directive for future engineering is the integration of diffusion distillation methodologies. Techniques such as Latent Consistency Models (LCMs) or Score Distillation Sampling could be applied

to both the LDM and the DiPT backbones. By training the network to directly predict the solution of the Probability Flow Ordinary Differential Equation (ODE) mapping noise to data, the generation process could be condensed from 1000 steps down to 1-4 steps. Evaluating the inevitable trade-off between this extreme inference acceleration and the degradation of sharp architectural high-frequency details will be a critical area of study.

6.2.6 Bridging the Modality Gap to Parametric B-rep Modeling

Finally, while GeoPriorPC successfully synthesizes high-fidelity, unstructured point clouds that capture complex architectural manifolds, raw point clouds are fundamentally a transitional representation. The ultimate geometric format required by BIM, CAD software, and physical simulation engines is a parametric, B-rep.

Future extensions of this pipeline must investigate appending a third, fully differentiable stage dedicated to structural reverse engineering. Utilizing the dense, uniform point clouds generated by GeoPriorPC as reliable geometric scaffolds, future networks could employ algorithms like RANSAC (Random Sample Consensus) or implicit neural representations (such as Neural Signed Distance Functions) to automatically segment planar surfaces and intersect them. By predicting an explicit graph of vertices, edges, and faces directly conditioned on our 3D diffusion output, the research community can fully close the pipeline, transforming raw, flat aerial pixels into fully parameterized, simulation-ready Urban Digital Twins.

Bibliography

[1] S. Turki, D. Panangian, H. Chaabouni-Chouayakh, and K. Bittner, "Aim2pc: Aerial image to 3d building point cloud reconstruction," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, pp. 1477–1484, 2025.

[2] F. Biljecki, J. Stoter, H. Ledoux, S. Zlatanova, and A. cCöltekin, "Applications of 3d city models: State of the art review," *ISPRS International Journal of Geo-Information*, vol. 4, no. 4, pp. 2842–2889, 2015.

[3] J. Argota Sánchez-Vaquerizo, "Urban digital twins and metaverses towards city multiplicities: uniting or dividing urban experiences?," *Ethics and Information Technology*, vol. 27, no. 4, 2025.

[4] A. Gonzalez-Caceres, F. Hunger, J. Forssén, S. Somanath, A. Mark, V. Naserentin, and A. Hollberg, "Towards digital twinning for multi-domain simulation workflows in urban design: a case study in gothenburg," *Journal of Building Performance Simulation*, vol. 18, no. 3, pp. 311–332, 2025.

[5] L. Wang, J. Li, W. Cen, and W. Feng, "Governing sustainable smart cities supported by city information modelling: a bibliometric analysis and systematic review," *Journal of Asian Architecture and Building Engineering*, vol. 24, no. 5, pp. 4433–4451, 2025.

[6] M. Deng, C. Menassa, and V. Kamat, "From bim to digital twins: a systematic review of the evolution of intelligent building representations in the aec-fm industry," *Journal of Information Technology in Construction (ITcon)*, vol. 26, pp. 58–83, 2021.

[7] X. Ye, S. Li, G. Gao, D. Retchless, Z. Cai, G. Newman, J. Du, and N. Duffield, "3d visualization of hurricane storm surge impact on urban infrastructure," *Urban Informatics*, vol. 3, no. 1, 2024.

[8] J. Huang, J. Stoter, R. Peters, and L. Nan, "City3d: Large-scale building reconstruction from airborne lidar point clouds," *Remote Sensing*, vol. 14, no. 9, p. 2254, 2022.

[9] R. Shao, J. Wu, J. Li, S. Peng, H. Chen, and C. Du, "Singlerecon: Reconstructing building 3-d models of lod1 from a single off-nadir remote sensing image," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 19588–19600, 2024.

[10] Y. Liu, L. Zhu, H. Ye, S. Huang, X. Gao, X. Zheng, and S. Shen, "Bwformer: Building wireframe reconstruction from airborne lidar point cloud with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[11] Y. Liu, A. Obukhov, J. D. Wegner, and K. Schindler, " Point2CAD: Reverse Engineering CAD Models from 3D Point Clouds ," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 3763–3772, IEEE Computer Society, June 2024.

[12] S. Huang, R. Wang, B. Guo, and H. Yang, "Pbwr: Parametric-building-wireframe reconstruction from aerial lidar point clouds," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27778–27787, 2024.

[13] M. A. Uy, Y. yu Chang, M. Sung, P. Goel, J. Lambourne, T. Birdal, and L. Guibas, "Point2cyl: Reverse engineering 3d objects from point clouds to extrusion cylinders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

[15] G. Zhan, C. Zheng, W. Xie, and A. Zisserman, "A general protocol to probe large vision models for 3d physical understanding," 2024.

[16] B. Ke, K. Qu, T. Wang, N. Metzger, S. Huang, B. Li, A. Obukhov, and K. Schindler, "Marigold: Affordable adaptation of diffusion-based image generators for image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[17] M. Bastico, D. Ryckelynck, L. Corté, Y. Tillier, and E. Decencière, "Rethinking metrics and diffusion architecture for 3d point cloud generation," 2025.

[18] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, "Point transformer v3: Simpler, faster, stronger," 2024.

[19] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2008.

[20] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[21] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[22] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 International Conference on 3D Vision (3DV)*, pp. 239–248, IEEE, 2016.

[23] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 5173–5176, IEEE, 2017.

[24] L. Mou and X. X. Zhu, "Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 143, pp. 1–14, 2018.

[25] P. Ghamisi and N. Yokoya, "Img2dsm: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 794–798, 2018.

[26] R. Peters, B. Dukai, S. Vitalis, P. van Oosterom, and J. Stoter, "Automated 3d reconstruction of lod2 and lod1 models for all 10 million buildings of the netherlands," *Photogrammetric Engineering & Remote Sensing*, vol. 88, no. 3, pp. 165–170, 2021.

[27] Y. Liu, R. Wang, S. Huang, and G. Cai, "Edgediff: Edge-aware diffusion network for building reconstruction from point clouds," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17008–17018, 2025.

[28] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

[29] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 523–540, Springer, 2020.

[30] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

[31] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987.

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

[33] Y. Liu, A. Obukhov, J. D. Wegner, and K. Schindler, "Point2building: Reconstructing buildings from airborne lidar point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 215, pp. 351–368, 2024.

[34] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6087–6101, 2021.

[35] T. Shen, J. Munkberg, J. Hasselgren, K. Yin, Z. Wang, W. Chen, Z. Gojcic, S. Fidler, N. Sharp, and J. Gao, "Flexible isosurface extraction for gradient-based mesh optimization," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–16, 2023.

[36] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, K. Kreis, and S. Fidler, "Get3d: A generative model of high quality 3d textured shapes learned from images," in *Advances In Neural Information Processing Systems*, vol. 35, pp. 31841–31854, 2022.

[37] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 300–309, 2023.

[38] Y. Siddiqui, A. Alliegro, A. Artemov, T. Tommasi, D. Sirin, M. Nießner, and J. Hou, "Meshgpt: Generating triangle meshes with decoder-only transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[39] Y. Chen, T. He, D. Huang, W. Ye, S. Chen, J. Tang, X. Chen, Z. Cai, L. Yang, G. Yu, G. Lin, and C. Zhang, "Meshanything: Artist-created mesh generation with autoregressive transformers," 2024.

[40] Y. Chen, Y. Wang, Y. Luo, Z. Wang, Z. Chen, J. Zhu, C. Zhang, and G. Lin, "Meshanything v2: Artist-created mesh generation with adjacent mesh tokenization," 2024.

[41] T. Hua, H. Gong, Y. Liu, D. Wang, Y.-C. Chen, and W. Zhao, "Buildanypoint: 3d building structured abstraction from diverse point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. Accepted, camera-ready version pending.

[42] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2837–2845, 2021.

[43] I. Corley and P. Najafirad, "Single-view height estimation with conditional diffusion probabilistic models," 2023.

[44] Z. Xiong, S. Chen, Y. Shi, and X. X. Zhu, "Self-supervised pretraining with monocular height estimation for semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[45] M. Viola, K. Qu, N. Metzger, B. Ke, A. Becker, K. Schindler, and A. Obukhov, "Marigold-dc: Zero-shot monocular depth completion with guided diffusion," 2025.

[46] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

[47] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in neural information processing systems*, vol. 30, 2017.

[48] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 5, pp. 1–12, 2019.

[49] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3075–3084, 2019.

[50] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel cnn for efficient 3d deep learning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[51] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

[52] P.-S. Wang, "Octformer: Octree-based transformers for 3d point clouds," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–11, 2023.

[53] Q. Hu, B. Yang, L. f. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11108–11117, 2020.

[54] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9224–9232, 2018.

[55] X. Lai, J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi, and J. Jia, "Stratified transformer for 3d point cloud segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8500–8509, 2022.

[56] A. Wichmann, A. Agoub, and M. Kada, "Roofn3d: Deep learning training data for 3d building reconstruction," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 1191–1198, 2018.

[57] R. Wang, S. Huang, and H. Yang, "Building3d: An urban-scale dataset and benchmarks for learning roof structures from point clouds," 2023.

[58] S. Huang, R. Wang, and X. Wang, "Buildingworld: A structured 3d building dataset for urban foundation models," 2025.

[59] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, U. Breitkopf, *et al.*, "The isprs benchmark on urban object detection and 3d building reconstruction," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, no. 3, pp. 293–298, 2012.

[60] B. Le Saux, N. Yokoya, R. Hänsch, and M. Brown, "2019 ieee grss data fusion contest: Large-scale semantic 3d reconstruction," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3297–3300, IEEE, 2019.

[61] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 253–256, 2018.

[62] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham, "Sensaturban: Learning semantics from urban-scale photogrammetric point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7154–7164, 2021.

[63] S. M. I. Zolanvari, S. Ruano, A. Rana, A. Cummins, R. E. da Silva, M. Rahbar, and A. Smolic, "Dublincity: Annotated lidar point cloud and its applications," 2019.

[64] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, vol. 27, 2014.

[65] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[66] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, pp. 2256–2265, PMLR, 2015.

[67] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[68] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020.

[69] X. Zeng, A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, and K. Kreis, "Lion: Latent point diffusion models for 3d shape generation," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 10021–10039, 2022.

[70] X. Ren, J. Huang, X. Zeng, K. Museth, S. Fidler, and F. Williams, "Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[71] Z. Lyu, B. Fei, J. Wang, X. Xu, Y. Zhang, W. Yang, and B. Dai, "Getmesh: A controllable model for high-quality mesh generation and manipulation," 2024.

[72] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv preprint arXiv:2212.08751*, 2022.

[73] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, "Emergent correspondence from image diffusion," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[74] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023.

[75] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv*, 2022.

[76] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[77] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[78] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 4296–4304, 2024.

[79] X. Fu, W. Yin, M. Hu, K. Wang, Y. Ma, P. Tan, S. Shen, D. Lin, and X. Long, "Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image," 2024.

[80] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, and W. Wang, "Wonder3d: Single image to 3d using cross-domain diffusion," 2023.

[81] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *arXiv preprint arXiv:1811.00982*, 2018.

[82] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402*, 2022.

[83] Y. Di, C. Zhang, P. Wang, G. Zhai, R. Zhang, F. Manhardt, B. Busam, X. Ji, and F. Tombari, "Ccd-3dr: Consistent conditioning in diffusion for single-image 3d reconstruction," 2023.

[84] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.

[85] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006.

[86] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, "Accelerating 3d deep learning with pytorch3d," *arXiv preprint arXiv:2007.08501*, 2020.

[87] L. Melas-Kyriazi, C. Rupprecht, and A. Vedaldi, "Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction," in *Arxiv*, 2023.