

GeoPriorPC: Nadir-view to 3D Point Cloud Reconstruction for buildings via Two-Stage Diffusion Priors

Youssef Korny¹, Sunghwan Yoo¹, Daniel Panangian², Ksenia Bittner²,
Andreas Wichmann³, Gunho Sohn¹

¹York University, Canada

²German Aerospace Center (DLR), Germany

³Jade University of Applied Sciences, Germany

{ykorny, jacobyyoo, gsohn}@yorku.ca

{daniel.panangian, ksenia.bittner}@dlr.de, andreas.wichmann@jade-hs.de

Abstract

Reconstructing 3D building point clouds from single nadir aerial images is a challenging task due to the inherent out-of-plane ambiguity. Existing diffusion-based methods rely on explicit 2D spatial features, which often struggle to accurately capture intricate architectural details, precise facade angles, and correct building proportions from top-down views. We propose **GeOPRIORPC**, a two-stage diffusion framework that leverages a learned latent geometric prior to guide 3D point cloud generation. In the first stage, we fine-tune a pre-trained latent diffusion model to predict oblique surface normals from nadir inputs. Operating directly in the compressed latent space of the normal map provides robust geometric features while avoiding the computational overhead of pixel-space decoding. In the second stage, a Diffusion Point Transformer (DiPT) reconstructs the 3D geometry using a dual-conditioning mechanism. This backbone integrates the latent normal prior globally via adaptive Layer Normalization (adaLN-Zero) to ensure stable training convergence, and locally through a cross-attention for fine-grained spatial alignment. Compared to state-of-the-art baselines, **GeoPriorPC** achieves a 44% relative improvement in *F1-score* and reduces the *Earth Mover’s Distance* by nearly 49%, demonstrating a clear capacity to generate accurate and uniform 3D building structures. Code can be found here: <https://github.com/youssefshaban/GeoPriorPC>

1. Introduction

The generation of 3D urban geometry from single-view aerial imagery represents a significant challenge in remote sensing, with critical applications in disaster response, flood

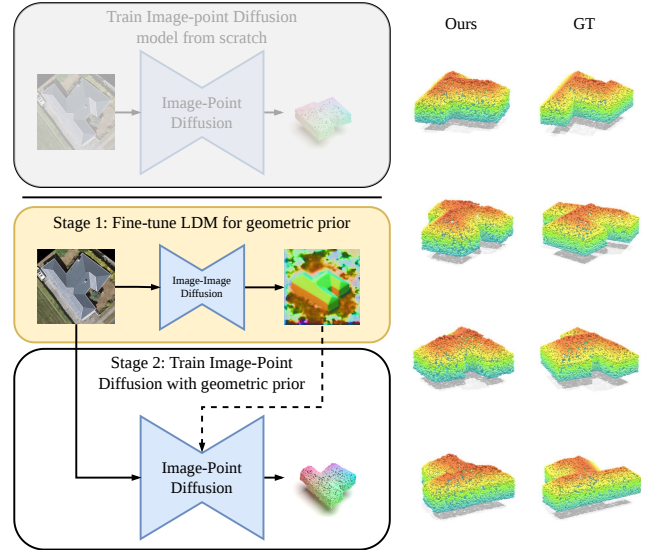


Figure 1. **Overview of the two-stage image-to-point diffusion framework.** The left panel illustrates our pipeline: a baseline trained from scratch (top) is compared to our approach, which uses a fine-tuned LDM to extract geometric priors (Stage 1) to guide 3D point cloud generation (Stage 2). The right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

modeling, and urban planning [2, 21, 28]. While high-fidelity 3D models can be generated using airborne LiDAR or multi-view photogrammetry [8, 12, 15], these acquisition methods are often prohibitively expensive for global-scale coverage. In contrast, nadir (top-down) satellite imagery is abundant and globally accessible. However, reconstructing accurate 3D geometry from a single nadir view is fundamentally ill-posed; the orthogonal projection captures roofs but occludes vertical geometry, requiring models to infer the

missing vertical dimension.

Traditional deep learning approaches often treat this as a regression problem, predicting Digital Surface Models (DSMs) or semantic height fields [11, 19]. Similarly, many large-scale reconstruction efforts focus on generating Level-of-Detail 1 (LoD1) models by extruding building footprints based on estimated heights [22, 23, 33]. While efficient for city-scale mapping, these prismatic representations fail to capture complex roof structures and true 3D manifolds. In contrast, point cloud generation offers a more flexible representation, yet existing methods such as AIM2PC [24] typically rely on lightweight backbones like PVCNN [13]. These architectures often struggle to fuse observed data with the plausible structural priors needed to reconstruct vertical geometry. We attribute this limitation to the lack of access to the rich, generalized world knowledge embedded in Large Foundational Models (LFMs) such as Stable Diffusion (SD) [20], which is a latent diffusion model (LDM). Recent interpretability studies demonstrate that pre-trained SD models encode linear representations of 3D depth, surface normals, and object support relations even without explicit 3D training [3, 29], while other methods achieve state-of-the-art in depth estimation and normal maps estimation by finetuning it [10].

To address this, we introduce *GeoPriorPC*, a novel two-stage framework that adapts these generative priors for 3D building reconstruction. Our approach decomposes the Image-to-Point Cloud problem into:

1. **Geometric Prior Learning (Image-to-Image):** We leverage a pre-trained LDM ecosystem (adapting the Marigold architecture [10]) to predict the surface normals of the corresponding oblique view. Unlike traditional methods that focus on RGB texture, our approach targets stable and sharp geometric maps, providing a robust shape prior that resolves nadir ambiguity.
2. **Geometry-Conditioned 3D Generation (Image-to-Point):** To exploit these priors, we adopt an architecturally diffusion point transformer (DiPT) backbone [1], which is inspired by the original point transformer v3 model (PTv3) [26]. We introduce a cross-attention mechanism that explicitly conditions the 3D denoising process on the generated 2D normal maps, allowing the model to "attend" to specific geometric cues while generating corresponding 3D points.

In summary, our contributions are:

- **Geometric Prior Transfer:** We present a novel integration pipeline that transfers the geometric knowledge of a pretrained LDM into a 3D point cloud diffusion model. Rather than using foundational models for 2D geometric inference alone, we retain the predicted latent normal representation and inject it into a 3D denoiser via a dual-conditioning mechanism, bridging 2D foundational priors and 3D generative reconstruction for the first time in the

geospatial domain.

- **Architectural Advancement:** We demonstrate that Transformer-based diffusion backbones (PTv3) offer significant performance gains over traditional CNN-based approaches (PVCNN) for geospatial reconstruction tasks.
- **Robust Conditioning Strategy:** We employ a cross-attention framework to seamlessly integrate 2D surface normal priors into the 3D generation pipeline, yielding state-of-the-art structural accuracy on the AIM2PC benchmark.

2. Related Work

2.1. Single-View 3D Reconstruction in Remote Sensing

Recovering 3D urban geometry from a single nadir-view image is an ill-posed problem that necessitates overcoming the inherent loss of the vertical dimension. Early deep learning frameworks approached this task via image-to-image translation, mapping RGB inputs to 2.5D Digital Surface Models (DSMs) using traditional encoder-decoder architectures [11, 19]. However, these representations are fundamentally constrained by their regular grid structure; by assigning a single z -elevation value to each (x, y) coordinate, they cannot represent co-planar vertical points, such as building facades, or complex structural undercuts. To scale reconstruction, recent pipelines decouple the problem into 2D instance segmentation for footprints followed by scalar height regression, producing Level-of-Detail 1 (LoD1) prismatic models [22, 23, 33]. While computationally efficient, this approach enforces a strict, often unrealistic assumption of perfectly flat roofs and simple vertical extrusion.

Generative point cloud methods, such as AIM2PC [24], bypass these representational limits by formulating the task as a conditional Denoising Diffusion Probabilistic Model (DDPM). AIM2PC projects 2D image features into 3D space using estimated camera poses and conditions its generation on low-level 2D representations like Sobel edge maps. In contrast, *GeoPriorPC* bypasses the limitations of low-level geometric cues by explicitly conditioning the 3D generation process on high-level, semantic geometric priors extracted from large-scale foundational models.

2.2. Generative Priors for 3D Reconstruction

To resolve the ambiguity inherent in single-view reconstruction, research has shifted from training task-specific networks from scratch to leveraging the rich latent knowledge within Large Foundational Models (LFMs). Interpretability studies indicate that text-to-image Latent Diffusion Models (LDMs) [20] implicitly encode a linear understanding of physical properties, including depth and surface normals, within their high-dimensional latent space [3, 29]. Extracting this geometry typically follows two paradigms. The first

utilizes Score Distillation Sampling (SDS) to optimize 3D representations using the 2D diffusion prior as a supervisory signal [16]. While structurally accurate, this iterative optimization process is computationally intractable for large-scale geospatial applications.

The second paradigm involves the feed-forward fine-tuning of the LDM. While architectures like ControlNet [30] introduced spatial conditioning, the Marigold framework [10] established a direct approach by fine-tuning the SD U-Net specifically for zero-shot depth estimation. This method effectively retains the high-frequency detail and semantic richness of the pre-trained VAE. Subsequent extensions, such as GeoWizard [7] and StableNormal [27], adapted these pipelines to generate deterministic surface normals. Our work closes the gap by utilizing the predicted oblique normal prior as an explicit intermediate representation, integrated into the point cloud diffusion process via global modulation and spatially-aware cross-attention to resolve missing vertical structures.

2.3. Deep Learning Architectures for Point Clouds

The choice of backbone architecture is critical for effectively processing the sparse and irregular nature of 3D point clouds. Early point-based networks like PointNet++ [17] utilized shared Multi-Layer Perceptrons (MLPs) and hierarchical pooling, but their reliance on localized radius searches often struggled to capture complex structural semantics over large scales. To improve efficiency, voxel-based methods like the Minkowski Engine [4] and Point-Voxel CNN (PVCNN) [13] introduced sparse 3D convolutions. While voxelization reduces memory consumption, it inherently introduces quantization loss and limited receptive fields—a significant bottleneck when synthesizing continuous vertical facades from a top-down view.

To overcome these spatial limitations, Transformer-based architectures have emerged as a scalable alternative. Initial designs like Point Transformer (PTv1) [31] improved feature aggregation via self-attention but were computationally bottlenecked by precise K-Nearest Neighbor (KNN) searches. The recent Point Transformer v3 (PTv3) [26] resolves this by replacing KNN with point cloud serialization, mapping unstructured points into structured 1D sequences. This enables efficient patch-based attention and expands the receptive field from 16 to 1024 points. In GeoPriorPC, we leverage this evolution by adopting PTv3 within our diffusion framework; its sequence-based processing naturally facilitates a robust cross-attention mechanism that globally correlates 2D normal maps with the expansive 3D point space.

3. Methods

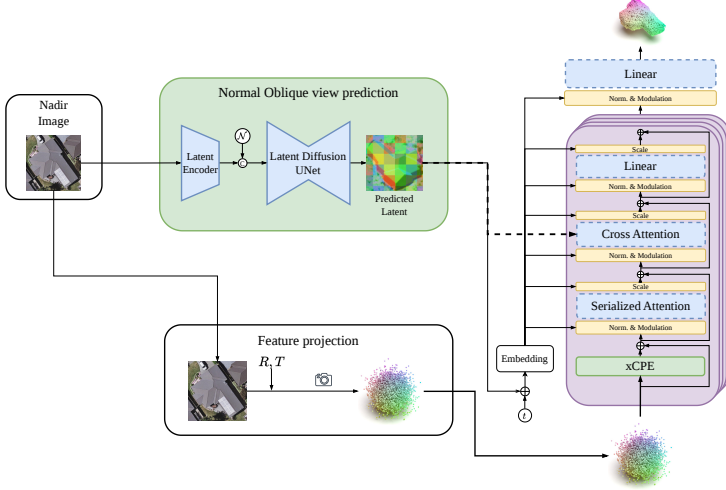
The architecture of our proposed *GeoPriorPC* framework is illustrated in Figure 2a. To resolve the ambiguity of recon-

structing vertical geometry from a single top-down view, we decouple the Image-to-Point Cloud task into a two-stage generative pipeline. As shown in Figure 2a (right), we first adapt Marigold to fine-tune a Latent Diffusion Model, translating nadir RGB images into structured oblique surface normals. Rather than decoding this output, we retain it as a compressed latent geometric prior. Figure 2b (left) depicts our complete generation pipeline. Nadir image features are projected onto a noisy 3D point cloud using known camera parameters. This point cloud is then iteratively denoised using a Point Transformer v3 (PTv3) diffusion backbone. Crucially, a cross-attention mechanism within the PTv3 blocks explicitly conditions 3D generation on the predicted latent normal map, fusing foundational 2D priors with 3D structural synthesis.

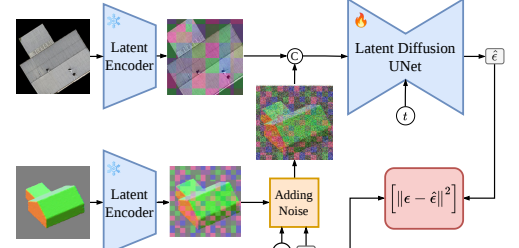
3.1. Preliminaries

Problem Formulation. Given a single nadir-view RGB aerial image $I \in \mathbb{R}^{H \times W \times 3}$, a corresponding building footprint mask $M_b \in \{0, 1\}^{H \times W \times 1}$, a Sobel edge map $E \in \mathbb{R}^{H \times W \times 1}$, and associated camera parameters $C = \{K, R, \mathbf{t}\}$. Because our dataset consists of standardized aerial acquisitions, we assume the intrinsic matrix K and translation vector $\mathbf{t}$ remain constant across all samples, leaving the rotation matrix R as the primary spatial variable. Our goal is to reconstruct a 3D building point cloud $P \in \mathbb{R}^{N_p \times 3}$, where N_p is the number of points. Note that P is defined in a normalized coordinate space; we focus on recovering the relative 3D structure and geometric proportions rather than absolute metric dimensions, which is consistent with existing generative benchmarks. While I , M_b , and E provide foundational 2D spatial and boundary context [24], they are insufficient for determining out-of-plane geometry. To resolve the inherent vertical ambiguity, our framework predicts an intermediate geometric prior to explicitly condition the 3D generation. Crucially, rather than operating in high-resolution pixel space, this prior is defined as a compressed latent surface normal representation $\hat{z}_0 \in \mathbb{R}^{h \times w \times c}$, where $h = H/8$, $w = W/8$, and $c = 4$ represent the downsampled latent dimensions.

Latent Diffusion Models (2D). To generate the intermediate normal map, we leverage Latent Diffusion Models (LDMs). Given a target image x , a pre-trained encoder $\mathcal{E}$ compresses it into a latent representation $z_0 = \mathcal{E}(x)$. The forward diffusion process progressively adds Gaussian noise to z_0 over T timesteps, producing a noisy latent z_t that approaches pure noise. The reverse generative process is then modeled as a Markov chain that iteratively removes this noise, mapping a random Gaussian vector $z_T \sim \mathcal{N}(0, \mathbf{I})$ back to a clean structured latent z_0 . To achieve this, a neural network ϵ_θ is trained to predict the added noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ given a conditioning signal c , op-



(a) Complete generation pipeline of GeoPriorPC. The 3D denoising process (DiPT [1]) uses cross-attention to explicitly condition on the latent normal map predicted from the nadir image.



(b) Geometric prior learning. The Marigold-based latent diffusion U-Net is fine-tuned using paired nadir images and rendered oblique normal maps.

Figure 2. **The proposed GeoPriorPC framework.** We decouple the problem into two stages: ((b)) fine-tuning an LDM to predict oblique latent normals from nadir inputs, and ((a)) a geometry-conditioned 3D diffusion process that maps 2D spatial features and structural priors to the 3D point space.

timized via the standard objective:

$$L_{LDM} = \mathbb{E}_{z_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2] \quad (1)$$

To reduce computations and preserve high-quality semantic features, we don't decode the output $\hat{z}_0$, but instead, we use it directly as our intermediate condition.

Point Cloud Diffusion (3D). For the final 3D generation stage, the diffusion process is extended to unstructured point sets. Let $P_0 = P$ denote the ground-truth 3D coordinates. The forward process perturbs these coordinates according to a variance schedule β_t :

$$q(P_t|P_{t-1}) = \mathcal{N}(P_t; \sqrt{1 - \beta_t}P_{t-1}, \beta_t\mathbf{I}) \quad (2)$$

The reverse denoising process iteratively recovers the clean point cloud P_0 from a standard Gaussian distribution $P_T \sim \mathcal{N}(0, \mathbf{I})$, conditioned on the input image I , mask M_b , edge map E , and our generated latent geometric prior $\hat{z}_0$.

3.2. Geometric Prior Learning

To resolve the out-of-plane ambiguity inherent in nadir aerial imagery, the first stage of our pipeline learns a dense structural prior by predicting oblique surface normals. We choose an oblique rendering viewpoint rather than a nadir-aligned one because nadir-view normals collapse vertical facade geometry onto a near-uniform plane, providing little discriminative signal for reconstructing out-of-plane structure. By contrast, oblique normals explicitly encode facade

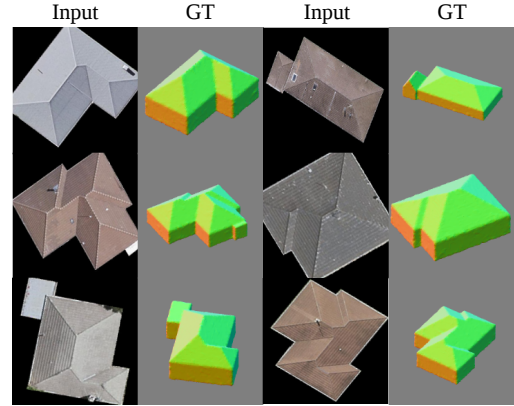


Figure 3. Samples from training dataset for geometric prior training. It shows pairs of aerial images as input and rendered normal maps as target.

orientation and roof slope angles, giving the 3D denoiser a richer geometric anchor from which to infer the missing vertical dimension. As illustrated in the right panel of Figure 2b, we adapt the Marigold framework [10], which fine-tunes the U-Net of a pre-trained Latent Diffusion Model (LDM) to extract robust geometric features.

Training Data Generation. Supervised fine-tuning of the LDM requires paired tuples of nadir RGB images and their corresponding ground-truth normal maps (I, N_{GT}) , as shown in Figure 3. Because the foundational build-

ing dataset provides only raw 3D point clouds, we derive N_{GT} programmatically. We first reconstruct dense 3D surface meshes from the ground-truth point clouds, and subsequently render the explicit geometric normals using a fixed camera pose C_N relative to the main camera pose C . This guarantees that all rendered normal maps maintain a consistent spatial relationship with the aerial image I , removing the need to provide C_N as an explicit condition for the model.

Latent Prior Learning and Extraction. To train the geometric prior, we map both the input aerial image I and the target normal map N_{GT} into the compressed latent space using the frozen VAE encoder $\mathcal{E}$, producing latents $z_I = \mathcal{E}(I)$ and $z_0 = \mathcal{E}(N_{GT})$, each of size $h \times w \times 4$. During training, Gaussian noise ϵ is progressively added to z_0 to generate a noisy latent z_t at timestep t . 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 $h \times w \times 8$. The diffusion U-Net ϵ_θ processes this concatenated input to predict the added noise ϵ , outputting a tensor of size $h \times w \times 4$. The optimization objective is formulated as:

$$L_{prior} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0, \mathbf{I}), t} [\|\epsilon - \epsilon_\theta(z_t \oplus z_I, t)\|_2^2] \quad (3)$$

where $\oplus$ denotes concatenation. During inference, we initialize a completely noisy target latent $z_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoise it over time using the fine-tuned U-Net, concatenating it with z_I at each step. Rather than passing the final denoised latent $\hat{z}_0$ through the VAE decoder, we use it directly as the intermediate spatial condition. Bypassing the decoder provides a compact, feature-rich representation that is easier for the main 3D generative model to learn from, while simultaneously reducing computational overhead.

3.3. Geometry-Conditioned 3D Generation

Multi-Modal Conditioning. The 3D generation process is guided by both explicit 2D spatial features and our learned latent geometric prior. Before each reverse diffusion step, we initialize a noisy point cloud P_t . For a given 3D coordinate $p_i \in P_t$, we project it onto the 2D image plane using the camera parameters C to obtain its continuous pixel coordinate $u_i = \Pi(p_i, C)$. We employ differentiable bilinear interpolation to accurately sample from the discrete 2D feature grids. The initial point-wise feature x_i is formed by concatenating these sampled explicit inputs—the RGB image I , the building footprint mask M_b , and the Sobel edge map E :

$$x_i = I(u_i) \oplus M_b(u_i) \oplus E(u_i) \quad (4)$$

Simultaneously, to inject the out-of-plane geometric structure, we process the compressed normal map $\hat{z}_0$ gener-

ated in Stage 1. Operating directly in the latent space preserves structural semantics and avoids pixel-space decoding overhead. A lightweight CNN encoder Φ_{local} extracts dense local features $F_{local} = \Phi_{local}(\hat{z}_0) \in \mathbb{R}^{h \times w \times d}$, which are retained for local cross-attention. To compute the global modulation parameters, a subsequent CNN module Φ_{seq} down-samples these features. A spatial average pooling operation followed by a linear projection matrix W_g compresses them into a condensed global context vector F_{global} :

$$F_{global} = W_g \cdot \text{AvgPool}(\Phi_{seq}(F_{local})) \in \mathbb{R}^d \quad (5)$$

Dual-Conditioned DiPT Backbone. We employ a Diffusion Point Transformer (DiPT) built on Point Transformer v3 (PTv3), which serializes sparse 3D point clouds into 1D sequences via space-filling curves for efficient self-attention. We integrate the learned geometric prior through two complementary pathways: global modulation and local feature matching.

First, the latent normal map is aggregated into a global descriptor F_{global} and fused with the timestep embedding e_t . An MLP computes scale and shift parameters, $[\gamma, \beta] = \text{MLP}(F_{global} + e_t)$, which modulate intermediate features X via adaptive Layer Normalization: $\text{adaLN}(X) = \gamma \odot \text{LayerNorm}(X) + \beta$. We employ an adaLN-Zero initialization strategy so the final MLP layer starts at zero; this ensures blocks initially act as identity mappings, preventing the geometric prior from destabilizing early training.

Second, for local feature matching, we flatten the latent prior into 2D tokens $\tilde{F}_{local} \in \mathbb{R}^{hw \times d}$. Following the self-attention layer (yielding 3D features X_{SA}), we apply a cross-attention mechanism. Here, the 3D point features act as queries to attend to the 2D normal tokens, which serve as both keys and values. To preserve spatial structure, the keys are augmented with 2D cosine positional embeddings PE . Finally, the output of this cross-attention block is projected via a weight matrix W_O and added back to the main feature stream through a residual connection:

$$X_{out} = X_{SA} + \text{CrossAttn}(Q, K, V)W_O \quad (6)$$

This dual-conditioning design explicitly bridges the 2D geometric prior with the 3D space, preserving both global structural proportions and local architectural details.

4. Experiments

4.1. Experimental Setup

Dataset and Evaluation Split. We evaluate on the AIM2PC dataset [24], containing paired nadir images, footprint masks, Sobel edge maps, and 3D point clouds. Lacking a canonical split for its 2,566 samples, we partitioned 2,248 for training and 318 for testing. For fair comparison, all baselines (e.g., AIM2PC [24], PC2 [14]) were retrained

from scratch on this split. Each ground-truth building comprises $N_p = 10,000$ points, globally normalized into a unit sphere for stable diffusion.

Geometric Prior Training (Stage 1). To prevent data leakage, we generate target normal maps using solely the 2,248 training point clouds. We estimate point normals via Open3D [32], reconstruct dense surfaces using Poisson Surface Reconstruction [9], and render oblique normal maps on-the-fly with PyTorch3D [18]. The LDM U-Net, initialized with Stable Diffusion 2.0 (Manojb/stable-diffusion-2-base), is fine-tuned following the standard Marigold [10] recipe, with inputs resized to 256×256 .

GeoPriorPC Implementation Details (Stage 2). 2D inputs are resized to 244×244 . To prevent coordinate drift, we adopt the Centered Denoising Probabilistic Model (CDPM) [24] recipe, maintaining a zero-mean point cloud across forward and reverse processes. For geometric prior encoding, CNN Φ_{local} outputs a $64 \times 64 \times d$ intermediate spatial resolution, pooling to a global context dimension $d = 256$. 3D generation models are trained for 100,000 steps (batch size 12) via AdamW on a single NVIDIA L40S GPU. The max learning rate is 1×10^{-3} , utilizing a Cosine Annealing scheduler (2,000 warmup steps) for the baseline and OneCycleLR for DiPT-based architectures [26].

Inference and Sampling Details. During test-time iterative reverse diffusion, Stage 1 samples the oblique normal latent $\hat{z}_0$ from Gaussian noise $z_T \sim \mathcal{N}(0, I)$ via DDIM ($T_1 = 20$ steps, linear noise schedule, $\eta = 0$, no classifier-free guidance). Stage 2 generates the point cloud from $P_T \sim \mathcal{N}(0, I)$ using DDIM ($T_2 = 1000$ steps), preserving CDPM zero-mean centering. We report single-sample generations with a fixed random seed of 60 for deterministic evaluation.

Evaluation Metrics. We evaluate 3D reconstruction quality using three unitless geometric metrics in normalized coordinate space: (1) Earth Mover’s Distance (EMD) [6], which measures transformation cost to penalize global structural anomalies; (2) Density-Aware Chamfer Distance (DCD) [25], which penalizes uneven point distributions to ensure surface uniformity; and (3) F1-score at a 0.02 threshold (F1@0.02) to balance spatial precision and recall. For the 2D normal map prediction, we calculate pixel-wise angular deviation using Mean Absolute Error (MAE) in degrees and the Sub 11.25° metric (the percentage of pixels with an error strictly below 11.25°).

4.2. Comparison with State-of-the-Art

We evaluate the proposed GeoPriorPC against three recent diffusion-based single-image 3D reconstruction baselines:

PC2 [14], CCD-3DR [5], and AIM2PC [24]. As shown in Table 1, our framework significantly outperforms all existing methods across every geometric metric.

Model	F1@0.02 ($\uparrow$)	EMD ($\downarrow$)	DCD ($\downarrow$)
PC2 [14]	0.484	0.00310	0.730
CCD-3DR [5]	0.512	0.00280	0.711
AIM2PC [24]	0.524	0.00255	0.696
GeoPriorPC (Ours)	0.754	0.00131	0.598

Table 1. Quantitative comparison of single-image 3D building reconstruction on the AIM2PC dataset. Best results are highlighted in **bold**.

The primary limitation of existing baselines is their exclusive reliance on explicit 2D spatial features. While methods like AIM2PC successfully capture 2D footprints by projecting RGB and Sobel edge features, they fundamentally struggle to resolve the inherent vertical ambiguity of nadir views. As illustrated in Figure 4, this often results in flattened or structurally distorted out-of-plane geometry.

By contrast, GeoPriorPC leverages a learned latent geometric prior to explicitly guide the 3D generation, successfully reconstructing precise facade angles and complex roof structures (Figure 4). This structural anchoring yields clear quantitative improvements: our method increases the F1-score by 44% relative to the previous state-of-the-art (AIM2PC) while reducing the Earth Mover’s Distance (EMD) by nearly half. Furthermore, the reduction in Density-Aware Chamfer Distance (DCD) confirms that the dual-conditioning mechanism within our PTV3 backbone not only captures the correct global shape but also generates highly uniform, artifact-free architectural surfaces.

4.3. Complexity and Inference Efficiency

We evaluate the model complexity and inference time of our proposed framework compared to the baseline. All timing evaluations are conducted on a single NVIDIA L40S GPU. For both transformer-based models (GeoPriorPC and GeoPriorPC-Concat), we maintain a consistent architecture depth of 6 blocks and a hidden dimension of $d = 256$.

As shown in Table 2, the full GeoPriorPC model contains 27.56M parameters and has an average iteration time of 0.154 seconds. Compared to the PVCNN-based AIM2PC baseline (28.05M parameters, 0.160 s/iter), our model requires fewer parameters and yields a lower iteration time. The GeoPriorPC-Concat variant contains the highest number of parameters (29.73M), but yields the lowest iteration time (0.122 s/iter) because it omits the local cross-attention computation. Overall, the full GeoPriorPC model achieves the highest geometric accuracy while remaining computationally lighter than the baseline backbone.

We also report the total end-to-end inference time per

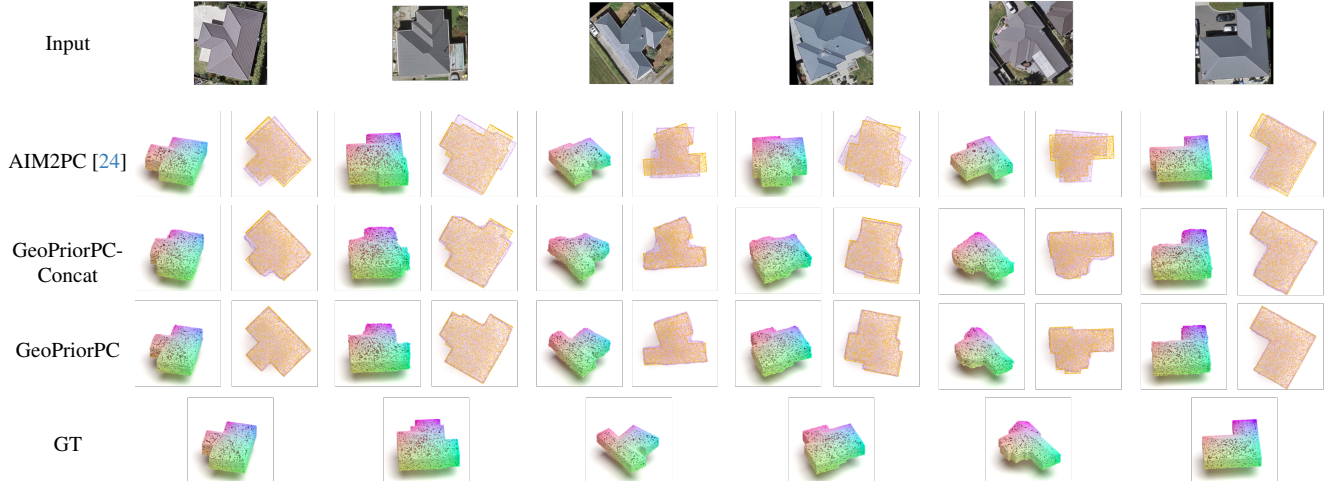


Figure 4. Qualitative comparison of 3D building reconstructions. From top to bottom: Input nadir image, AIM2PC baseline, our model with global context only, our full GeoPriorPC model, and Ground Truth.

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

Table 2. Comparison of model complexity and average iteration time evaluated on an NVIDIA L40S GPU.

sample, which combines Stage 1 DDIM sampling ($T_1 = 20$ steps at 1.5 s/iter) and Stage 2 point cloud generation ($T_2 = 1000$ steps at 0.154 s/iter). On a single NVIDIA L40S GPU, the full pipeline requires approximately 184 seconds per sample, compared to 160 seconds for the AIM2PC baseline. Despite the two-stage design, the total inference overhead is very close between them.

4.4. Ablation Studies

We conduct an ablation study to evaluate the individual contributions of our chosen backbone and conditioning strategies. The results are summarized in Table 3.

First, we isolate the effect of the backbone architecture. Upgrading from the baseline AIM2PC model to the DiPT architecture without normal conditioning (GeoPriorPC-Ignore) improves the F1-score from 0.524 to 0.565. This indicates that the PTv3-based backbone processes point cloud features more effectively than the baseline.

Second, we evaluate the impact of the geometric prior. Injecting the global context of the normal map via simple concatenation (Concat) yields improvements for both the baseline (F1: 0.560) and the DiPT backbone (F1: 0.678). However, the global context alone lacks fine-grained spatial alignment. Our final model, which utilizes both global

adaLN-zero injection and local cross-attention, achieves the best results across all metrics. This demonstrates that local feature matching between the 3D points and the 2D prior tokens is necessary for accurate structural reconstruction.

Model Variation	F1@0.02 ($\uparrow$)	EMD ($\downarrow$)	DCD ($\downarrow$)
<i>Baseline Backbone (AIM2PC)</i>			
Baseline (No Prior)	0.524	0.00255	0.696
+ Global Context (Concat)	0.560	0.00220	0.682
<i>Proposed Backbone (PTv3)</i>			
GeoPriorPC-Ignore (No Prior)	0.565	0.00216	0.683
GeoPriorPC-Concat (Global Only)	0.678	0.00187	0.633
GeoPriorPC (Global + Local)	0.754	0.00131	0.598

Table 3. Ablation study on architectural backbones and prior conditioning strategies.

4.5. Impact of Pre-trained Latent Priors

To validate the necessity of using pre-trained LDM weights for the geometric prior, we compare our fine-tuned Marigold model against an identical architecture initialized from scratch. The quality of the predicted oblique normal maps is evaluated using Mean Absolute Error (MAE) and the percentage of pixels with an angular error below 11.25° .

As shown in Table 4, the model trained from scratch fails to learn an accurate mapping from nadir images to out-of-plane normals, resulting in a high MAE of 49.5° and capturing only 5.5% of pixels within the acceptable error threshold. This numerical failure is visually evident in Figure 6, which illustrates how the outputs from the randomly initialized model collapse and fail to capture meaningful geometric structure. In contrast, the fine-tuned SD model achieves an MAE of 19.3° and a Sub 11.25° accuracy of 58.3%. As demonstrated in Figure 5, the fine-tuned model accurately

resolves complex roof structures and vertical planes. This indicates that the pre-trained SD weights are required to successfully resolve the structural ambiguities inherent in nadir aerial imagery.

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

Table 4. Evaluation of normal map prediction quality comparing random initialization versus fine-tuning pre-trained LDM weights.

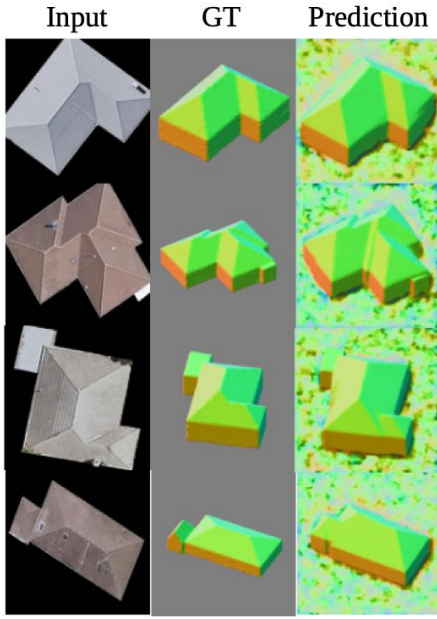


Figure 5. Qualitative results from fine-tuned LDM after geometric prior training.

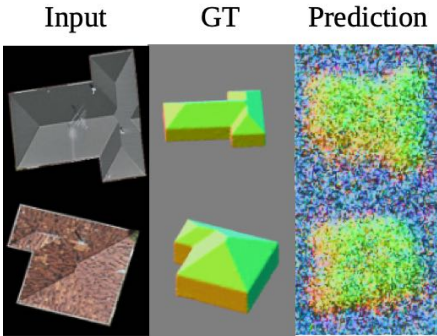


Figure 6. Qualitative results from an LDM model trained from scratch on the same dataset. It shows that the model collapsed while training.

5. Limitations and Future Work

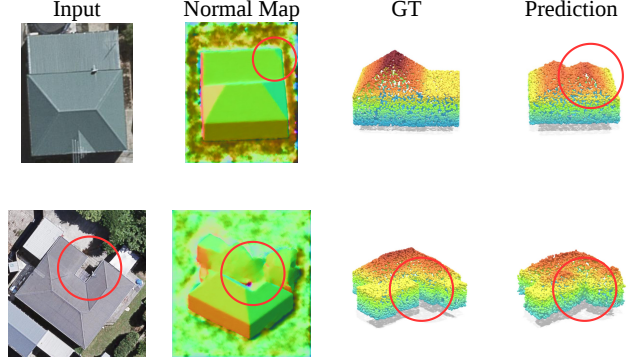


Figure 7. Failure cases demonstrating error propagation. Geometric inconsistencies in the LDM-predicted normal maps lead to corresponding structural artifacts in the final 3D point clouds.

While *GeoPriorPC* effectively resolves out-of-plane ambiguities from single nadir-view inputs, its generalization is currently bounded by the limited scale and architectural diversity of the AIM2PC training dataset. Furthermore, as a sequential framework, the final 3D point cloud is highly dependent on the accuracy of the intermediate latent representations. Specifically, if the Latent Diffusion Model (LDM) predicts an incorrect or physically inconsistent normal map, the subsequent generation stage inherits and propagates these geometric errors into the final structure, as illustrated in Figure 7. Additionally, this two-stage diffusion process introduces high inference error inherited from the normal maps. Future work will focus on training with larger, more diverse urban datasets to scale the model’s capacity and improve the robustness of the latent prior against intermediate prediction failures. We also plan to explore diffusion distillation techniques to accelerate inference for efficient, city-scale 3D building reconstruction.

6. Conclusion

This paper presented *GeoPriorPC*, a two-stage diffusion framework for reconstructing 3D building point clouds from single nadir images. To address the lack of vertical cues, we introduced a learned latent geometric prior based on oblique surface normals. This prior is integrated into a Diffusion Point Transformer (DiPT) backbone using both global adaLN-Zero injection and local cross-attention. Our approach outperforms existing baselines on the AIM2PC dataset, achieving a 44% relative improvement in F1-score and a 49% reduction in Earth Mover’s Distance, demonstrating a strong capacity to generate accurate and structurally uniform 3D buildings.

References

- [1] Matteo Bastico, David Ryckelynck, Laurent Corté, Yannick Tillier, and Etienne Decencière. Rethinking metrics and diffusion architecture for 3d point cloud generation. *arXiv preprint arXiv:2511.05308*, 2025. 2, 4
- [2] Filip Biljecki, Jantien Stoter, Hugo Ledoux, Sisi Zlatanova, and Arzu Çöltekin. Applications of 3d city models: State of the art review. *ISPRS International Journal of Geo-Information*, 4(4):2842–2889, 2015. 1
- [3] Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720*, 2023. 2
- [4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
- [5] Yan Di, Chenyangguang Zhang, Pengyuan Wang, Guangyao Zhai, Ruida Zhang, Fabian Manhardt, Benjamin Busam, Xiangyang Ji, and Federico Tombari. Ccd-3dr: Consistent conditioning in diffusion for single-image 3d reconstruction, 2023. 6
- [6] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 6
- [7] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. 2024. 3
- [8] Jin Huang, Jantien Stoter, Ravi Peters, and Liangliang Nan. City3d: Large-scale building reconstruction from airborne lidar point clouds. *Remote Sensing*, 14(9):2254, 2022. 1
- [9] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 6
- [10] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis, 2025. 2, 3, 4, 6
- [11] Chao-Jung Liu, Vladimir A. Krylov, Paul Kane, Geraldine Kavanagh, and Rozenn Dahyot. Im2elevation: Building height estimation from single-view aerial imagery, 2020. 2
- [12] Yujia Liu, Anton Obukhov, Jan Dirk Wegner, and Konrad Schindler. Point2building: Reconstructing buildings from airborne lidar point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215:351–368, 2024. 1
- [13] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. In *Advances in Neural Information Processing Systems*, 2019. 2, 3
- [14] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12834–12844, 2023. 5, 6
- [15] Massimiliano Pepe, Domenica Costantino, Vincenzo Saverio Alfio, and Ariana Grimaldi. An effective method for 3d modelling of urban areas from aerial images and lidar data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W6-2024: 25–30, 2024. 1
- [16] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 3
- [17] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3
- [18] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 6
- [19] Yoonas Rezaei and Stephen Lee. sat2pc: Estimating point cloud of building roofs from 2d satellite images, 2022. 2
- [20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2
- [21] Ary Setijadi Prihatmanto, Abdurrakhman Prasetyadi, Ambar Yoganingrum, Ridwan Sutriadi, and Ana Hadiana. The digital twin city in enhancing flood evacuation systems: Future opportunities and challenges. *IEEE Access*, 13:32371–32383, 2025. 1
- [22] Ruizhe Shao, JiangJiang Wu, Jun Li, Shuang Peng, Hao Chen, and Chun 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*, 17:19588–19600, 2024. 2
- [23] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. Continental-scale building detection from high resolution satellite imagery, 2021. 2
- [24] 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*, XLVIII-G-2025:1477–1484, 2025. 2, 3, 5, 6, 7
- [25] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. In *Advances in Neural Information Processing Systems*, pages 23696–23706, 2021. 6
- [26] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger, 2024. 2, 3, 6
- [27] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2024. 3

- [28] Xinyue Ye, Shoujia Li, Ge Gao, David Retchless, Zhenhang Cai, Galen Newman, Jiaxin Du, and Nick Duffield. 3d visualization of hurricane storm surge impact on urban infrastructure. *Urban Informatics*, 3(1), 2024. [1](#)
- [29] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding, 2024. [2](#)
- [30] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. [3](#)
- [31] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. [3](#)
- [32] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. [6](#)
- [33] Xiao Xiang Zhu, Sining Chen, Fahong Zhang, Yilei Shi, and Yuanyuan Wang. Globalbuildingatlas: An open global and complete dataset of building polygons, heights and lod1 3d models. *Earth System Science Data*, 17(12):6647–6668, 2025. [2](#)