Appearance-aware Scaling Diffusion Model for 3D Point Cloud Upsampling

Sunghwan Yoo[1], Gunho Sohn[1]*

[1] Dept. of Earth and Space Science and Engineering, York University, Toronto, ON, M3J 1P3 Canada - (jacobyoo, gsohn)@yorku.ca

Keywords: Airborne LiDAR, Point Cloud Upsampling, Diffusion Models, Scene Reconstruction, Deep Learning

Abstract

Airborne laser scanning (ALS) point clouds are widely used for large-scale 3D scene understanding, but acquiring dense ALS data remains costly and sparse observations often exhibit incomplete vertical structures and uneven sampling. Existing diffusion-based point cloud upsampling methods have shown promise, yet conditioning only on sparse point coordinates often leads to noisy surfaces, boundary artifacts, and limited structural recovery in under-observed regions. In this work, we propose the **Appearance-aware Scaling Diffusion Model (ASDM)**, a conditional diffusion framework for scene-level ALS point cloud upsampling that incorporates multi-view projected depth-image cues derived directly from the input point cloud. Specifically, sparse ALS scenes are rendered from multiple virtual viewpoints to generate projected depth images, which provide complementary structural information for guiding the denoising process. These projected-image features are fused with sparse point features to improve geometric fidelity and scene-level consistency. For training and evaluation, we construct realistic sparse–dense scene pairs from aerial LiDAR data derived from the YUTO Semantic dataset using a region-disjoint split over 13 survey areas. Experiments under the $\times 4$ upsampling setting show that ASDM outperforms recent diffusion-based baselines, achieving the best overall performance in Chamfer Distance (0.5643), JSD-3D (0.6688), F1-score (75.67), and voxelized IoU at 1 m and 2 m resolutions. These results demonstrate that projected-image conditioning is an effective strategy for robust airborne LiDAR scene densification.

1. Introduction

Airborne laser scanning (ALS) point clouds are a fundamental data source for large-scale 3D scene understanding and support applications such as urban digital twins, infrastructure monitoring, environmental mapping, and large-area surveying. Compared with terrestrial or mobile laser scanning, ALS enables efficient acquisition over broad geographic regions and is therefore widely used in remote sensing pipelines that require scalable 3D measurements. However, acquiring dense and accurate ALS point clouds remains costly, as higher point density often requires lower flight altitudes, slower flight speeds, or repeated overlapping flight paths, all of which increase survey time and operational expense (Lohani and Ghosh, 2017). In practice, ALS data also exhibit highly uneven sampling patterns due to viewpoint geometry, occlusion, and sensor trajectory. Horizontal surfaces such as roofs and ground regions are often relatively well observed, whereas vertical structures such as building facades, thin objects, and sharp scene boundaries remain sparse or incomplete. These characteristics can significantly degrade downstream tasks such as semantic segmentation, surface reconstruction, and 3D scene modeling, motivating the need for effective scene-level point cloud upsampling methods that can recover missing geometric detail from sparse airborne observations.

Traditional point cloud upsampling methods have primarily focused on object-centric or patch-level reconstruction, where the goal is to generate denser local point distributions from relatively clean and compact inputs. While these methods have shown promising results for local shape refinement, they are less well suited to large-scale airborne scenes, where sparsity is spatially non-uniform and structural incompleteness often occurs over extended regions. Scene-level ALS upsampling is particularly challenging because it requires preserving both local geometric detail and global structural coherence across complex urban layouts, vegetation, and mixed open-space environments. In under-observed regions, especially along vertical boundaries and sparse architectural surfaces, purely geometry-driven upsampling may struggle to infer reliable missing structure from point coordinates alone.

Recent advances in denoising diffusion probabilistic models (Ho et al., 2020) have demonstrated strong potential for 3D generative modeling (Luo and Hu, 2021) and point cloud upsampling (Nunes et al., 2024, Qu et al., 2024). Diffusion-based approaches are particularly attractive because they can model complex point distributions and iteratively reconstruct missing structure from sparse observations through a learned denoising process. Among recent methods, conditional diffusion frameworks have shown promising performance for sparse point cloud completion and LiDAR upsampling. Nevertheless, upsampling sparse ALS scenes remains difficult. When conditioned only on sparse point coordinates, existing methods often struggle to preserve structural fidelity, leading to noisy surfaces, boundary artifacts, and uneven point generation concentrated near already observed samples rather than under-observed regions. As a result, recovering sharp boundaries and geometrically coherent scene structure from sparse airborne scans remains an open challenge.

To address these limitations, we propose the **Appearance-aware Scaling Diffusion Model (ASDM)**, a conditional diffusion framework for scene-level ALS point cloud upsampling that incorporates multi-view projected-image cues derived directly from the input point cloud. Specifically, the sparse point cloud is rendered from multiple virtual viewpoints to produce projected depth images, which provide complementary structural information about surface layout, edge boundaries, and visibility patterns that may be difficult to infer from sparse 3D coordinates alone. These projected observations are encoded and

* Corresponding author

fused with sparse point features to guide the diffusion process toward more complete and structurally consistent reconstructions. By introducing appearance-aware multi-view conditioning, ASDM encourages the model to better recover missing structure in sparsely observed regions while preserving overall scene-level geometric consistency.

For training and evaluation, we construct sparse–dense point cloud pairs from aerial LiDAR scenes derived from the YUTO Semantic dataset (Yoo et al., 2023), enabling supervised scene-level upsampling under realistic airborne sensing conditions. The resulting benchmark follows a region-disjoint split to evaluate generalization on geographically unseen survey areas, providing a realistic setting for assessing scene-level airborne LiDAR densification.

The main contributions of this work are summarized as follows:

- We propose the **Appearance-aware Scaling Diffusion Model (ASDM)**, a conditional diffusion framework for scene-level ALS point cloud upsampling that enhances sparse-point conditioning with multi-view projected-image cues.
- We introduce a **projected-image conditioning strategy** that fuses sparse point features with multi-view depth projections, enabling the diffusion model to better recover structural details in under-sampled regions and along scene boundaries.
- We construct a **realistic sparse–dense training and evaluation protocol** from the YUTO Semantic dataset for supervised scene-level ALS upsampling under region-disjoint generalization.
- We demonstrate through experiments on real-world airborne LiDAR scenes that the proposed appearance-aware conditioning improves geometric fidelity and structural consistency compared with existing diffusion-based point cloud upsampling baselines.

2. Related Work

2.1 Point Cloud Upsampling and Scene Reconstruction

Point cloud upsampling aims to reconstruct a denser and geometrically faithful point set from sparse and irregular observations. Early learning-based approaches such as PU-Net established a representative framework for point cloud upsampling by learning dense point distributions from local geometric features (Yu et al., 2018). While many upsampling methods have shown strong performance on object-centric shapes and local patches, scene-level reconstruction remains more challenging due to large spatial extent, irregular sampling, and structural complexity.

More recently, point cloud upsampling has expanded toward LiDAR-oriented and scene-level settings. In particular, LiDiff and PUDM formulate sparse point cloud densification as a conditional generative process and demonstrate the potential of learned upsampling for outdoor LiDAR scenes (Nunes et al., 2024, Qu et al., 2024). However, most existing approaches are not specifically designed for ALS, where anisotropic point density, sparse facade observations, and large-scale urban structure make scene-level upsampling substantially more difficult. These characteristics motivate methods that can better preserve structural boundaries and recover missing detail from sparse ALS observations.

2.2 Diffusion Models for Point Cloud Processing

Denoising diffusion probabilistic models (DDPMs) have emerged as a powerful generative framework due to their stable optimization and strong distribution modeling capability (Ho et al., 2020). Their success has motivated growing interest in 3D point cloud generation and reconstruction, beginning with diffusion-based point cloud generation and extending to conditional restoration tasks such as upsampling and scene reconstruction (Luo and Hu, 2021, Nunes et al., 2024, Qu et al., 2024). These developments suggest that diffusion models are well suited for recovering missing structure and enriching sparse 3D observations.

Among existing methods, LiDiff and PUDM are most relevant to our setting because they explicitly formulate sparse point cloud upsampling as a conditional diffusion process (Nunes et al., 2024, Qu et al., 2024). However, their conditioning is primarily driven by sparse point observations alone, which can limit structural fidelity in severely under-sampled ALS scenes. In contrast, our proposed **Appearance-aware Scaling Diffusion Model (ASDM)** introduces multi-view projected-image conditioning to provide complementary structural cues for improved scene-level airborne LiDAR reconstruction.

3. Background

3.1 Denoising Diffusion for Conditional Point Cloud Generation

Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) are generative models that synthesize data by progressively transforming Gaussian noise into a target sample through an iterative denoising process. A DDPM consists of a forward diffusion process that gradually perturbs clean data with Gaussian noise and a learned reverse process that reconstructs the sample from noisy observations.

Given a clean point cloud $x_0 \sim p_{\text{data}}(x)$, the forward diffusion process is defined as

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

where $\beta_t \in (0,1)$ is the noise variance at timestep t. Let $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Then, the closed-form transition from x_0 to x_t is

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\mathbf{I}\right). \tag{2}$$

During training, a denoising network $\epsilon_\theta(x_t, t)$ is optimized to predict the injected Gaussian noise ϵ using the standard objective

$$\mathcal{L}_{\text{DDPM}} = E_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]. \tag{3}$$

For conditional generation, the reverse process is guided by an external condition c, such as a sparse point cloud or learned conditioning features. The reverse transition is formulated as

$$p_\theta(x_{t-1} \mid x_t, c) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, c, t), \sigma_t^2 \mathbf{I}\right), \tag{4}$$

and the corresponding conditional training objective becomes

$$\mathcal{L}_{\text{cond}} = E_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, c, t)\|_2^2\right]. \tag{5}$$

In point cloud upsampling, the condition c typically corresponds to the sparse input scene or its derived embeddings, guiding the reverse diffusion process toward dense and geometrically consistent reconstructions.

Among recent conditional diffusion methods for 3D point clouds, LiDiff (Nunes et al., 2024) is particularly relevant because it is designed for scene-level generation under sparse LiDAR observations. Unlike point-set generation formulations that rely heavily on local patch normalization or aggressive coordinate rescaling, LiDiff performs denoising directly in the original scene coordinate space, which helps preserve absolute spatial structure and local geometric relationships throughout the diffusion process. This property is especially important for ALS upsampling, where large-scale layout, point spacing, and structural boundaries must remain consistent across extended outdoor scenes. In addition, LiDiff employs a sparse convolutional denoising backbone built on the Minkowski Engine, which performs convolution directly on occupied voxels rather than dense volumetric grids. This allows efficient feature extraction from large-scale sparse point clouds while maintaining spatial locality and computational efficiency. For these reasons, we adopt LiDiff as the generative backbone in ASDM and focus on improving its conditional guidance rather than redesigning the diffusion process itself. Specifically, ASDM follows the LiDiff framework while augmenting the conditioning pathway with projected-image cues to provide complementary structural information for scene-level airborne LiDAR upsampling.

4. Methodology

4.1 Overview

Given a sparse airborne LiDAR scene patch $P_s \in R^{N\times 3}$, the goal is to generate a dense point cloud $P_d \in R^{rN\times 3}$ that preserves both local geometric detail and large-scale scene structure, where r denotes the upsampling ratio. Compared with object-level point cloud upsampling, scene-level ALS upsampling is more challenging due to anisotropic sampling, sparse facade observations, and large-scale urban structure.

To address these challenges, we propose the **Appearance-aware Scaling Diffusion Model (ASDM)**, a conditional diffusion framework that augments sparse point observations with complementary image-space structural cues. ASDM renders the sparse input point cloud from multiple virtual viewpoints to generate projected depth images, extracts projected-image features, and fuses them with sparse point features before injecting the result into the LiDiff denoising backbone (Nunes et al., 2024). An overview of the proposed framework is shown in Fig. 1.

4.2 Projected-Image Conditioning

Sparse 3D point coordinates alone often provide insufficient information in severely under-sampled ALS regions, particularly near occlusion boundaries and sharp structural transitions. To provide complementary structural cues, ASDM adopts a **RealisticProjection** module inspired by prior multi-view point cloud projection strategies (Mei et al., 2024). The input point cloud is rendered from ten virtual camera viewpoints to produce projected depth images together with visibility masks and projected pixel coordinates for feature reprojection.

Each projected depth image is processed by a pretrained CLIP ViT-B/32 visual encoder to extract dense image-space features. Although CLIP is pretrained on natural RGB imagery, rasterized depth projections preserve structural patterns such as boundaries, depth discontinuities, and surface layout, which can still provide useful shape-aware cues for feature extraction. For each 3D point, features from visible projections are aggregated across the ten views and passed through a lightweight multilayer perceptron to obtain a 256-dimensional projected-image embedding aligned with the original 3D coordinates.

This projected-image conditioning provides complementary cues that are difficult to infer from sparse coordinates alone, particularly in under-observed regions where point density is low or spatial support is incomplete. As a result, the diffusion model is better guided toward structurally consistent and spatially coherent reconstruction.

4.3 Feature Fusion

To integrate sparse point information with projected-image cues, ASDM employs a lightweight transformer-based fusion block, referred to as the **Cross-Modal Feature Fusion** module. Given a sparse-point embedding and a projected-image embedding for each point, both of dimension $D = 256$, the two modality features are organized as modality tokens $\mathbf{T} \in R^{N\times 2\times D}$.

A multi-head attention layer with 8 heads is applied along the modality dimension, enabling the sparse-point and projected-image tokens to exchange complementary information. The attended tokens are then processed by a feed-forward network with residual connections and layer normalization. Finally, the fused representation is obtained by mean aggregation across the modality dimension,

$$\tilde{f}_i = \frac{1}{2}\sum_{m=1}^{2} \mathbf{T}_{i,m}, \tag{6}$$

followed by a linear projection and ReLU activation to produce the final per-point conditioning feature $f_i \in R^{512}$.

The fused features are then organized as MinkowskiEngine SparseTensors so that the original point coordinates remain aligned with the sparse convolutional denoising backbone. This fusion strategy allows ASDM to jointly exploit sparse geometric observations and projected-image structural context in a unified conditioning representation.

4.4 Integration with the Diffusion Backbone

ASDM is implemented as a conditioning enhancement on top of the LiDiff diffusion backbone (Nunes et al., 2024), enabling a controlled comparison with sparse-point-only diffusion conditioning. Specifically, we retain the same denoising network architecture, diffusion noise schedule, and conditional noise-prediction objective as LiDiff, while replacing the original conditioning with the fused representation produced by the proposed projected-image and fusion modules.

Following the standard denoising diffusion probabilistic model (DDPM) formulation (Ho et al., 2020), Gaussian noise is progressively added to the target dense point cloud during the forward process, and the denoising network is trained to predict the injected noise at each timestep conditioned on the noisy sample and the fixed conditioning representation. During inference, the model follows the reverse denoising process to generate the final dense point cloud from noise.

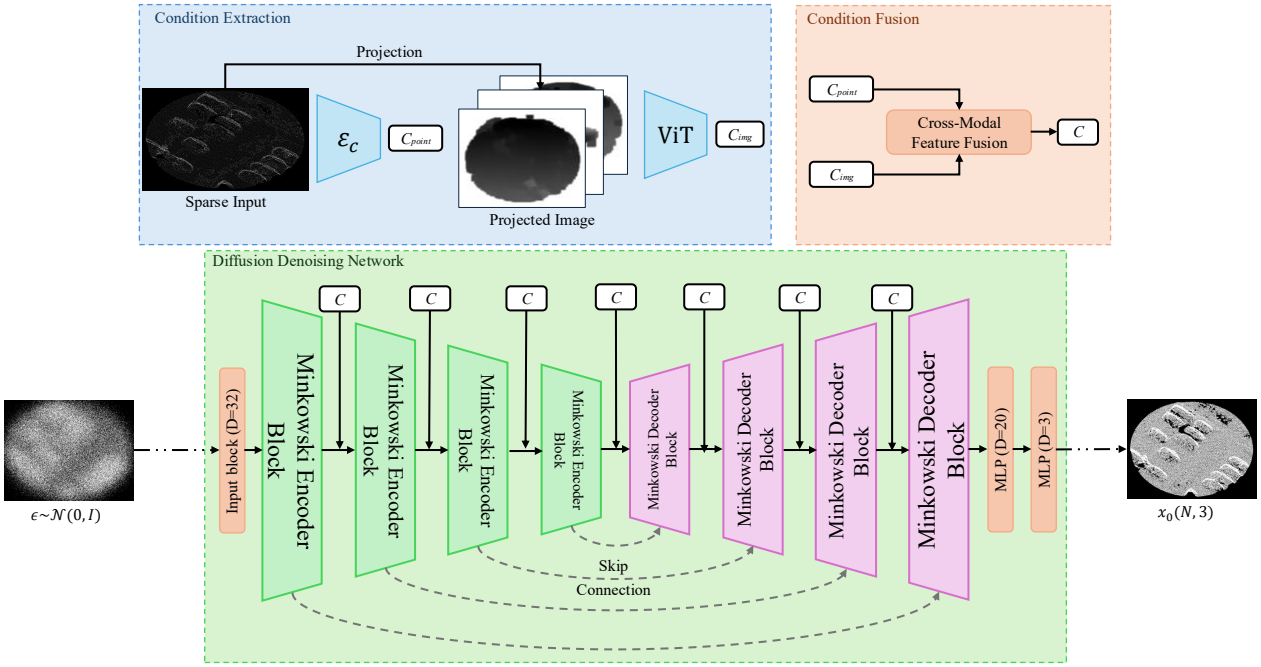

Figure 1. Overview of ASDM. Sparse ALS points are rendered into multi-view projected depth images, encoded into image-space features, and fused with sparse point features to form the conditioning representation for the diffusion backbone, which reconstructs a dense point cloud.

By preserving the underlying diffusion formulation and backbone design, performance gains can be attributed primarily to the improved conditioning signal rather than modifications to the generative model itself.

5. Experiments

5.1 Data Preparation

For training and evaluation, we construct realistic sparse–dense scene pairs from ALS data derived from the YUTO Semantic dataset (Yoo et al., 2023), which was collected over the York University campus. The source data consist of multiple overlapping ALS flight lines captured under real airborne acquisition conditions, providing natural variation in sampling density, scan geometry, and visibility across repeated observations. Unlike synthetic downsampling protocols that randomly discard points from a single dense scan, this setup enables sparse supervision that more closely reflects real-world airborne LiDAR sparsity caused by acquisition overlap, viewpoint differences, and uneven surface coverage.

To generate sparse–dense pairs, we exploit the *Scan ID* metadata associated with each flight line to separate the aggregated point cloud into its constituent strips. Overlapping regions between strips are then identified to establish corresponding sparse and dense observations. Within these overlapping areas, local scene samples are extracted by uniformly sampling centroids and collecting all points within a 30 m radius neighborhood around each centroid. To reduce redundancy, sampling is performed in a coverage-aware manner so that previously selected neighborhoods are not repeatedly reused. In regions containing more points than required, random uniform sampling is applied to enforce consistent point counts across samples. Each sparse input contains 12,000 points, while denser targets are constructed at $\times 2$, $\times 3$, and $\times 4$ scales, corresponding to 24,000, 36,000, and 48,000 points, respectively.

This process yields 1,300 localized scene samples from 13 survey areas. To ensure robust generalization and prevent spatial leakage, we adopt a region-disjoint split at the survey-area level, using 10 areas for training, 2 for validation, and 1 for testing. All results in this paper are reported on the challenging $\times 4$ upsampling setting. This protocol provides a realistic and reproducible benchmark for scene-level ALS point cloud upsampling under genuine airborne sensing variability.

5.2 Experimental Setup

All experiments are conducted using the region-disjoint train/validation/test split described in Section 5.1. Unless otherwise stated, all quantitative results are reported on the held-out test set. To focus the evaluation on a challenging setting, we conduct all experiments under the $\times 4$ upsampling protocol, where each sparse input scene is upsampled from 12,000 points to 48,000 points.

For quantitative comparison, all baseline methods are retrained using the same input–output protocol. ASDM is implemented on top of the LiDiff backbone (Nunes et al., 2024) and follows the same training configuration for a controlled comparison. We evaluate performance using complementary metrics that assess geometric fidelity, distributional consistency, local correspondence, and volumetric completeness. Specifically, geometric fidelity is measured using Chamfer Distance (CD), Hausdorff Distance (HD), and Earth Mover’s Distance (EMD). Distributional consistency is assessed using Jensen–Shannon Divergence (JSD) in both bird’s-eye-view (BEV) and 3D voxel space. Local correspondence is evaluated using the F1-score under distance thresholds of 0.5–1.0 m. Volumetric completeness is measured using voxelized Intersection-over-Un-ion (IoU) at resolutions of 0.5 m, 1.0 m, and 2.0 m.

Model	CD ↓	HD ↓	JSD BEV ↓	JSD 3D ↓	EMD ↓	F1 ↑	0.5 m IoU ↑	1 m IoU ↑	2 m IoU ↑
LiDiff (Nunes et al., 2024)	0.6057	8.9752	0.4392	0.7029	**3.0006**	72.4891	0.1436	0.3103	0.4552
PUDM (Qu et al., 2024)	1.2277	**6.5701**	0.4673	0.6958	7.1298	63.8475	**0.1862**	0.3338	0.4721
ASDM (Ours)	**0.5643**	8.6653	**0.4325**	**0.6688**	3.3695	**75.6653**	0.1798	**0.3476**	**0.4911**

Table 1. Quantitative comparison of scene-level ALS upsampling performance under the $\times 4$ setting (12,000 $\rightarrow$ 48,000 points). Lower is better for CD, HD, JSD, and EMD, while higher is better for F1 and voxelized IoU.

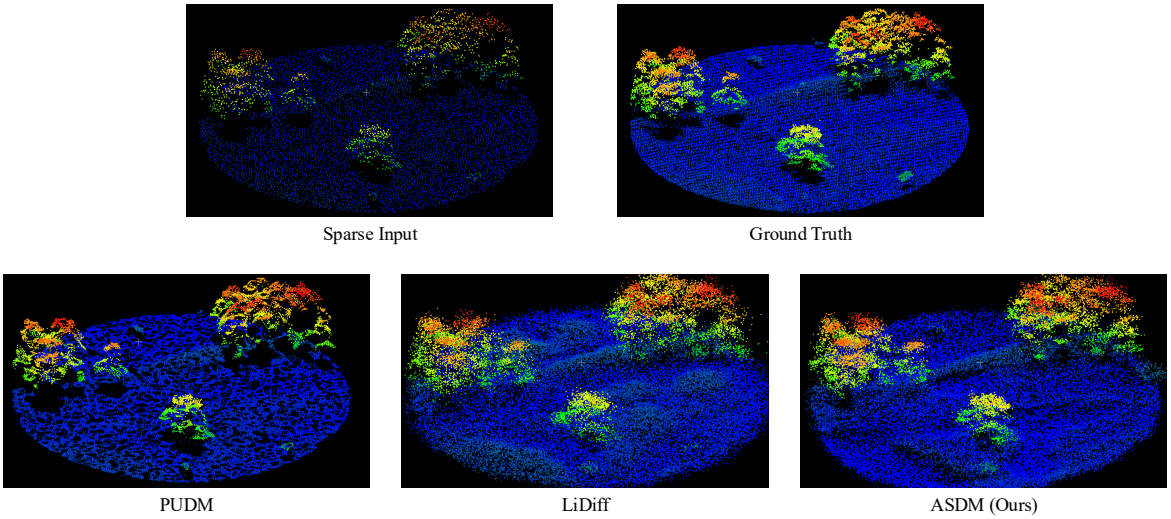

Figure 2. Qualitative comparison of scene-level ALS upsampling results under the $\times 4$ setting. ASDM produces denser and more structurally consistent reconstructions.

5.3 Quantitative Results

We compare ASDM against two recent diffusion-based point cloud upsampling baselines: LiDiff (Nunes et al., 2024), which serves as the backbone and sparse-point-only baseline, and PUDM (Qu et al., 2024), a recent conditional diffusion model for point cloud upsampling. For a fair comparison, all methods are retrained using the same sparse–dense pair construction and region-disjoint split described in Section 5.1. Quantitative results under the $\times 4$ upsampling setting are summarized in Table 1.

ASDM achieves the best overall performance across most evaluation metrics, including CD, JSD-BEV, JSD-3D, F1, and voxelized IoU at 1 m and 2 m resolutions. Compared with LiDiff, the proposed appearance-aware conditioning improves both geometric fidelity and structural consistency, indicating that projected-image cues provide useful complementary information beyond sparse point observations alone. In particular, ASDM reduces CD from 0.6057 to 0.5643 and improves F1 from 72.49 to 75.67, while also yielding more consistent gains in JSD and volumetric IoU.

PUDM achieves a lower Hausdorff Distance and slightly higher IoU at the finest 0.5 m voxel resolution. However, its overall performance is less stable, with substantially worse CD, EMD, JSD-BEV, and F1 compared with ASDM. This suggests that while PUDM may better recover isolated local occupancy in some fine-resolution regions, ASDM provides a stronger overall balance between local geometric accuracy and globally coherent scene reconstruction for sparse ALS upsampling.

Qualitative visual comparisons in Fig. 2 further support these findings, showing that ASDM produces denser and more structurally consistent reconstructions.

Model	CD ↓	JSD_{3D} ↓	F1 ↑
Concat fusion	0.7519	0.6903	76.8987
Concat + Mink. layer	0.6204	0.6749	**77.2823**
Cross-modal fusion	**0.5643**	**0.6688**	75.6653

Table 2. Comparison of fusion strategies for combining point and projected-image conditions on the region-disjoint test split under $\times 4$ upsampling.

Model	CD ↓	JSD_{3D} ↓	F1 ↑
Low-res projection	**0.5643**	**0.6688**	75.6653
High-res projection	0.6480	0.6713	**76.7169**

Table 3. Effect of projected-image resolution on scene-level ALS upsampling with ASDM.

5.4 Fusion Strategy Ablation

To examine how projected-image features should be integrated into the diffusion backbone, we compare three fusion strategies for combining point and projected-image conditions. As shown in Table 2, simple feature concatenation yields the weakest overall performance, indicating that naive fusion is insufficient for robust scene-level conditioning. Adding an additional Minkowski layer after concatenation substantially improves both CD and JSD_{3D}, while also achieving the highest F1 score. However, the proposed cross-modal fusion strategy achieves the best CD and JSD_{3D}, indicating stronger geometric fidelity and distributional consistency. Although its F1 score is slightly lower than the concatenation-based variant with an additional Minkowski layer, the overall results suggest that the proposed fusion design provides a better balance between local correspondence and global structural reconstruction in sparse ALS scenes.

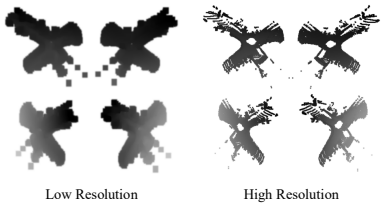

Figure 3. Qualitative comparison of low- and high-resolution projected-image conditioning in ASDM. The low-resolution setting generally produces more spatially continuous reconstructions, while the high-resolution setting may recover slightly finer local detail in some regions.

5.5 Effect of Projected-Image Resolution

To examine the effect of the projected-image conditioning scale, we compare low- and high-resolution projection settings within ASDM. As shown in Table 3, increasing the projection resolution does not consistently improve performance. Although the high-resolution setting yields a slightly higher F1 score, the low-resolution configuration achieves better Chamfer Distance (CD) and JSD in 3D voxel space. This suggests that lower-resolution projections provide a more stable conditioning signal by emphasizing global structural layout, whereas higher-resolution projections may capture finer local detail but introduce noisier correspondences in sparse ALS scenes. Qualitative examples in Fig. 3 support this observation, showing that the low-resolution setting generally produces more spatially continuous and structurally coherent reconstructions, particularly in sparse architectural regions.

6. Conclusion

In this work, we presented **ASDM**, a conditional diffusion framework for scene-level airborne LiDAR upsampling that incorporates projected-image conditioning to improve structural reconstruction from sparse ALS observations. Built on a realistic sparse–dense training protocol derived from the YUTO Semantic dataset, ASDM leverages complementary 2D structural cues to guide the reverse diffusion process toward denser and more geometrically consistent scene-level point cloud reconstructions. Experimental results demonstrate that the proposed conditioning strategy improves overall reconstruction quality over existing diffusion-based baselines, highlighting the value of projected-image guidance for scene-level ALS upsampling.

In future work, we plan to expand and publicly release the current benchmark to support broader research in realistic airborne LiDAR upsampling across more diverse scenes, land-cover conditions, and acquisition settings. We also aim to investigate additional conditioning signals beyond projected images, including other geometric and multi-modal cues, to further improve robustness across different sensing environments. In particular, we plan to extend the experimental evaluation to other LiDAR acquisition modalities such as mobile laser scanning (MMS) and terrestrial laser scanning (TLS), where occlusion patterns, viewpoint anisotropy, and sampling characteristics differ substantially from airborne scans. Beyond geometric densification, the proposed conditioning framework may also be extended to related tasks such as void filling, partial scene completion, and attribute-aware point generation, including intensity or color prediction. Exploring joint geometric and radiometric upsampling therefore remains an important direction for future work. In addition, while ASDM is built on the LiDiff backbone in this work, future research will investigate alternative and more conventional diffusion formulations to better understand how different denoising architectures affect scene-level point cloud generation under varying LiDAR acquisition conditions. We hope this work encourages future research in condition-aware point cloud densification and generative scene reconstruction for real-world remote sensing applications.

References

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840–6851.

Lohani, B., Ghosh, S., 2017. Airborne LiDAR technology: A review of data collection and processing systems. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, 87(4), 567–579.

Luo, S., Hu, W., 2021. Diffusion probabilistic models for 3d point cloud generation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2837–2845.

Mei, G., Riz, L., Wang, Y., Poiesi, F., 2024. Geometrically-driven aggregation for zero-shot 3d point cloud understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nunes, L., Marcuzzi, R., Mersch, B., Behley, J., Stachniss, C., 2024. Scaling diffusion models to real-world 3d lidar scene completion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14770–14780.

Qu, W., Shao, Y., Meng, L., Huang, X., Xiao, L., 2024. A conditional denoising diffusion probabilistic model for point cloud upsampling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20786–20795.

Yoo, S., Ko, C., Sohn, G., Lee, H., 2023. Yuto Semantic: a Large Scale Aerial LIDAR Dataset for Semantic Segmentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 209–215.

Yu, L., Li, X., Fu, C.-W., Cohen-Or, D., Heng, P.-A., 2018. Pu-net: Point cloud upsampling network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2790–2799.